Study of the oral virome and microbiome associated to the proliferative verrucous leukoplakia

Doctorando: Rodrigo García López

Programa: 1393/2007 Biotecnología,

FACULTAT DE CIÈNCIES BIOLÒGIQUES

Directores: Andrés Moya Simarro

Vicente Pérez Brocal

Valencia, 2017

D. Andrés Moya, Catedrático del Departamento de Genética de la Universitat de València y D. Vicente Pérez Brocal, Investigador del Centro de Investigación Biomédica en Red (Epidemiología y Salud Pública) (CIBERESP), certifican que la memoria titulada "**Study of the oral virome and microbiome associated to the proliferative verrucous leukoplakia**" ha sido realizada bajo su dirección en la Unidad Mixta de Investigación en Genómica y Salud FISABIO-Salud Pública -Instituto Cavanilles de Biodiversidad y Biologia Evolutiva, Universitat de València.

Y para que así conste, firman el presente certificado.

Fdo. Dr. Andrés Moya

Fdo. Dr. Vicente Pérez Brocal

Valencia, a ____

Agradecimientos

A mi esposa, Letty por el enorme apoyo y cariño en estos momentos tan complicados.

A Andrés por su interminable paciencia y la motivación para seguir adelante.

A Vicente y Ma. José por la valiosa guía tanto en lo académico como en la vida.

A Jorge y Bea por la amistad que tanta falta hace para poder disfrutar del trabajo.

A Maria por marcar la pauta y recordarme siempre hay algo positivo en todo.

A los compañeros del laboratorio por la convivencia y a los técnicos por su imprescindible ayuda.

A los que ya no están, los que nos dejaron pero a quienes recordamos con mucho cariño.

A mi papá, por los valiosos consejos que quisiera seguir más frecuentemente.

A mi mamá por ser quien me aterriza en la realidad cuando divago.

A mis hermanos, por recordarme que no todo en la vida es trabajo.

A La Bola, por recordarme que la amistad es para toda la vida.

A los amigos familiares y compañeros que me ayudaron a llegar a este punto y muchos nombres que no olvido pero que no hace falta mencionar para saber que los estimo y siento su apoyo.

A todos, mis más sinceras gracias.

"WE ARE ALL LIKELY TO MAKE MISTAKES, EXCEPT THOSE OF US WHO DO NOTHING."

C. BERNARD
AN INTRODUCTION TO THE STUDY OF EXPERIMENTAL MEDICINE, 1865

“THE DEVELOPMENT OF SOMETHING NEW, EVEN IN HUMAN ENDEAVOUR, IS GENERALLY THE OUTCOME OF EFFORTS MANY OF WHICH YIELD ONLY NEGATIVE RESULTS.”

A. J. KLUYVER AND C. B. VAN NIEL,
THE MICROBE'S CONTRIBUTION TO BIOLOGY, 1956

La realización de este trabajo fue apoyada por proyectos del Ministerio de Economía y Competitividad de España SAF 2012-31187, SAF2013-49788-EXP y SAF2015-65878-R, la beca de Formación de Personal Investigador del Ministerio de Ciencia e Innovación de España y la beca CONACYT-CECTI del Consejo Nacional de Ciencia y Tecnología de México y el Consejo Estatal de Ciencia y Tecnología de Michoacán.

TABLE OF CONTENTS

INTRODUCTION

General Introduction

It is a small world after all

Recent figures place the estimated number of prokaryotic cells on Earth between $9.2 \cdot 10^{29}$ and $31.7 \cdot 10^{29}$ (*1*), making them the most abundant living organisms on the planet, with insects being next at around 10^{19} (*2*). These numbers were put together from cell density calculations in samples taken from a wide variety of habitats, factorized by the total biomass of each habitat. As impressive as they may be, their biomass only accounts for around 8 to 32 % of the total sum that is believed to exists globally (*1*, *3*). Furthermore, the number of viral particles is allegedly ten times that of prokaryotes, based on a conservative estimate that has been extrapolated from observations from marine samples (*4*).

Available statistics have an ample margin of error but, even so, they help us put the actual composition of life on our planet into context. Microbes are abundant, as much as they are ubiquitous, colonizing much diverse ecosystems and undoubtedly playing a significant role in natural biogeochemical cycles (*5*). A 2016 report by Locey and Lennon predicted that the Earth may be inhabited by an overwhelming 10^{11}-10^{12} total microbial species (*6*), whereas eukaryotic species have been calculated to be in the order of 10^{7} (*7*) and that is not even considering viral species. Nevertheless, the vast majority of the microbial world remains unexplored since no suitable culturable techniques have been found for most organisms (*8*, *9*).

Towards a multi-omics era

The advent of metagenomics marked the dawn of the era of multidisciplinary meta-omics (*10*), fuelled by the appearance of the second generation of automated sequencers back in the mid-2000s (*11*). This has provided, over the course of over a decade, an ever-growing set of approaches to study the complex collection of microbes in environmental samples: the microbiota.

There exists no globally accepted consensus as to how broad and what exactly the terms “microbiome” and “microbiota” encompass. This is still a matter of debate as publications use them differently depending on the context (Box 1). For the sake of this document’s understanding the text will refer to microbiota as the complete collection of all microbes, living or not. That is, viruses and viroids will be included in the term, unless stated otherwise. The word microbiome will be avoided to prevent confusion, addressing the type of material and its source directly (e.g. bacterial/viral metagenome/metatranscriptome) but when used it will refer to the microscopic biomes. Metagenomics will refer to the science of unculturable genomics while metagenome will refer both to the whole genomic contents of the microbiota and to the techniques for studying it (whole genomic shotgun sequencing or WGS). Viromics will broadly refer to the study of all material related to viral origin, including Eukaryotic viruses, prokaryotic phages, integrated genomic elements such as prophages and endogenous viruses and viroids.

Box 1.- Important definitions

Most authors employ the term “microbiota”, a replacement for its now deprecated analogue “microflora”, to refer to the actual living microorganisms in an environmental setting. Some use “microbiome” to refer to all their corresponding genomic content and mobile elements associated to the microbiota (*374*), similarly to how the word “genome” refers to a single organism’s genetic material.

Instead, other articles use the term “microbiome” with a broader ecological definition, closer to that of the microbiota. They refer to all microscopic life found in a given environment (*375*). This is analogous to the way the word “biome” refers to macroscopic organisms in an ecosystem and the definition may as well include genomes and products derived from it (Figure 1). To make matters more complicated, some works also consider the viral fraction of environmental samples as part of the microbiome, despite viruses not actually being recognized as living organisms (*376*).

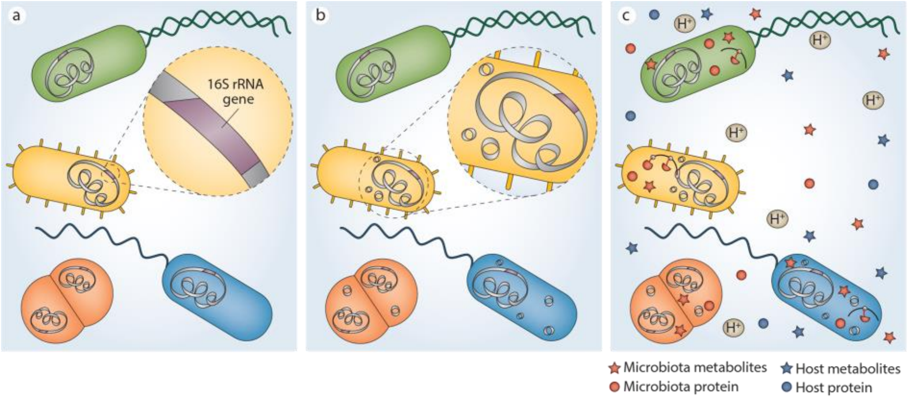


Figure 1 Different terms used to describe the microbiota and related analyses. A) Definition of the microbiota, metagenome and microbiome. A) Microbiota: All microorganisms present in the environment. Taxonomic diversity can be assessed with 16S profiling. B) Metagenome: Genes and genomes of the microbiota, including plasmids. Functional diversity can be assessed by whole genome sequencing. C) Microbiome: The complete microscopic ecological niche including the microbiota, the metagenome and products both from the microbiota and the environment. Multiple omics sciences study its different compartments. From: Whiteside et.al 2015 (*72*)

Regarding more technical terms, the word “metagenome” traditionally referred to the whole unbiased (i.e. non-amplified) genomic content that was extracted from an environmental sample but has been extensively employed to refer to any culture-independent collection of sequences such as amplicon-based 16S profiling used for taxonomic explorations.

Additionally, the concept of holobiont addresses the symbiotic (thus biont) ecological unit formed by a host and its associated microbiota as a whole (Greek “hólos”) (*377*). This supraorganism concept is relevant for higher order multicellular Eukaryotes which are very important natural reservoirs for the latter.

The development of metagenomics as a science is tightly linked to advances in research technologies. Back in 2005, for the first time since the inception of DNA sequencing, high-throughput platform GS20 by 454 Life Sciences had started producing thousands of sequences generating over 20 Mpb in a single execution, representing the first commercially competitive approach for genomic sequencing (*12*). This paved the way forward to a new revolution in the emerging field of metagenomic research (and the multiple other meta-omic approaches that would soon follow) (*13*, *14*).

Before this, few researchers had been able to personally get a glimpse at the actual functional and taxonomic diversity that existed in mixed microbial populations from environmental samples. The notion of the existence of an uncultivable fraction was there since the early 1930s, cued by the large discrepancy in the proportions of plate counts and direct counts (*15*). However, most estimates came from microscopic counts that relied on manually computing Gram-stained organisms (*16*).

The year 1977 marked a milestone in the history of microbiology. Frederick Sanger and collaborators had just published a refined method for sequencing DNA molecules using dideoxy chain terminators (*17*). Also, Carl Woese and George E. Fox proposed the first wide-spectrum phylogenetic analysis based upon a universal marker present in all life forms, the small subunit of the ribosomal RNA (16S/18S rRNA gene or rDNA) with the few sequences available at the time, effectively separating the Archaea and Bacteria domains in the tree of life for the first time (*18*).

Although it was not the only molecular marker in use, the rRNA in 16S profiling was endorsed for taxonomic classification by the microbiologist community (*19*). Still, the sequencing process was reliant on having adequate amounts of DNA molecules, which in turn depended on culture techniques. This all changed with the publication of Kary Mullis' polymerase chain reaction (PCR) method in 1988 (*20*), a technique that allowed the exponential amplification of fragments of DNA using flanking primers (short DNA oligos used to start amplification in complementary strands of the template).

In 1990, Stephen J. Giovannoni and collaborators published a novel sequencing experiment in which they amplified environmental bacterial DNA (no culturing involved) using a PCR targeted at the 16S rDNA of 12 randomly selected sequences in samples from the Sargasso Sea (*21*). They reported the existence of a new clade of bacteria, based on a cluster of phylogenetic divergent sequences that was called SAR11 and that it is now known to be formed by a non-culturable group of small, carbon-oxidizing bacteria that comprise around 25% of all plankton (*22*).

The rest of the 90s underwent an increase in the popularity of 16S profiling techniques, expanding the niches that were surveyed as it became clearer that non-culturable organisms were virtually ubiquitous and existing across much varied phyla, from novel cyanobacteria in the hot springs of the Yellowstone National Park (*23*) to new members of Proteobacteria and Firmicutes in industrial-contaminated sediment from near Seattle (*24*) and previously unknown Bacteroidetes in human faecal samples (*16*).

The Ribosomal Database Project was established by Bonnie Maidak and collaborators (among whom was Carl Woese) as an ftp mail server in 1994 (*25*). This was the first effort to make data readily available via an online repository providing curated alignments of rDNA sequences published until then. Despite this, it would not be until the end of the century that the term “metagenome” would be coined by Handelsman and collaborators to refer to the “collective genomes of soil microflora [sic]” (*26*).

By the turn of the new millennium, and borrowing from the early methods of genomic sequencing, the scientific community had finally managed to study some of the genetic contents of non-culturable organisms (other than rRNA genes) using bacterial artificial chromosomes (BACs, Figure 2A) and other DNA constructs as vectors for cloning fragments of their genomes obtained directly from environmental marine and soil samples (*27*, *28*). It was not long before groups reported the construction of the first metagenomic libraries that were processed by random shotgun sequencing (WGS, Figure 2B) and genomic assembling, first securing parts of viral genomes from uncultured marine communities in just a little more than a thousand sequences (*29*). Eventually, researches obtained near-complete prokaryotic genomes with sets over 70 Mbp from very specialized acid mine drainage biofilms (*30*) and even get a much more complete diversity estimation from massive clone sets of over 1.36 Gbp (giga [10^9] base pairs) from surface water samples in the Sargasso sea (*31*). All these experiments settled the basis of metagenomics analyses that would follow.

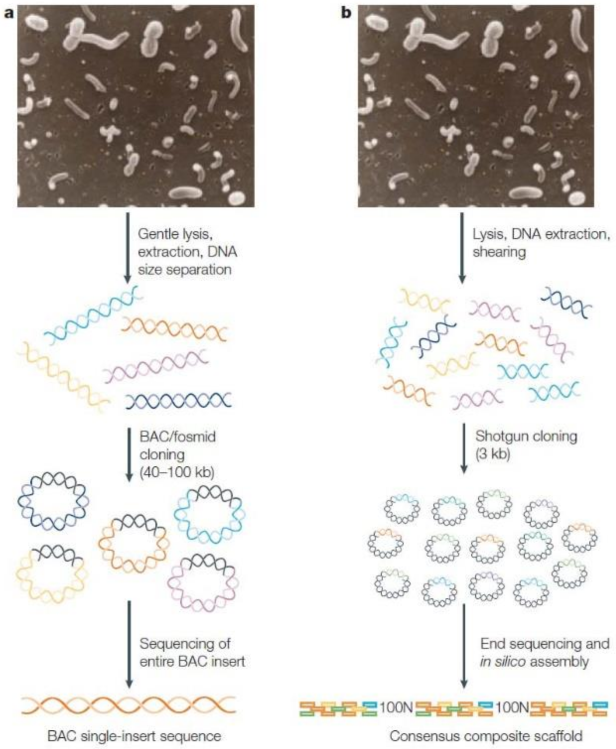

a
Gentle lysis, extraction, DNA size separation
BAC/fosmid cloning (40–100 kb)
Sequencing of entire BAC insert
BAC single-insert sequence
b
Lysis, DNA extraction, shearing
Shotgun cloning (3 kb)
End sequencing and *in silico* assembly
100N
100N
Consensus composite scaffold

Figure 2. Schematic diagram of early approaches for retrieving genomic sequence information from natural microbial populations. A) Large DNA inserts are recovered in bacterial artificial chromosomes (BACs) that are each derived from an individual cell. Subsequent assembly results in a contiguous DNA sequence derived from a single cell in the original population. B) Small inserts from a genetically heterogeneous population are cloned. Sequence assembly is derived from multiple different cells. From: DeLong 2005 (*32*).

Metagenomics did not take off smoothly, mainly due to the forbidding complexity of producing the multiple clones necessary to construct a thorough metagenomic library, coupled with the prohibitive costs of sequencing the genomic shotgun fragments with the Sanger method (*17*), even though the first generation of fully automated sequencing platforms was available, producing large reads (700-900bp) but at a very slow pace (< 80Kbp per day) (*33*).

The arrival of a new generation of sequencing platforms vastly changed the landscape of metagenomics. This second generation of automated sequencers was able to manage real-time sequencing using PCR-amplified fragments as templates.

Rise and fall of sequencing platforms

Originally intended for genomic sequencing, the *454 Genome Sequencer* system employed a sequencing-by-synthesis approach that used a pyrosequencing protocol (*34*). In it, more than a million of previously-sheared sequences were first attached to microbeads, each holding only a single sequence (Figure 3). They were then amplified individually via an emulsion PCR, producing millions of different clonal sets of molecules that were deposited in plates containing picolitre wells that could hold only one bead each. The attached sequences were used as templates for sequencing in a programmatic flood cycle of nucleotide incorporation that was then detected by the release of inorganic pyrophosphate and the subsequent generation of photons. A charge-coupled device (CCD) sensor captured light emissions from wells that incorporated nucleotides and the intensities were used to determine how many of them were incorporated in each cycle (in the case of homopolymers). Cycles were repeated until sequencing was done.

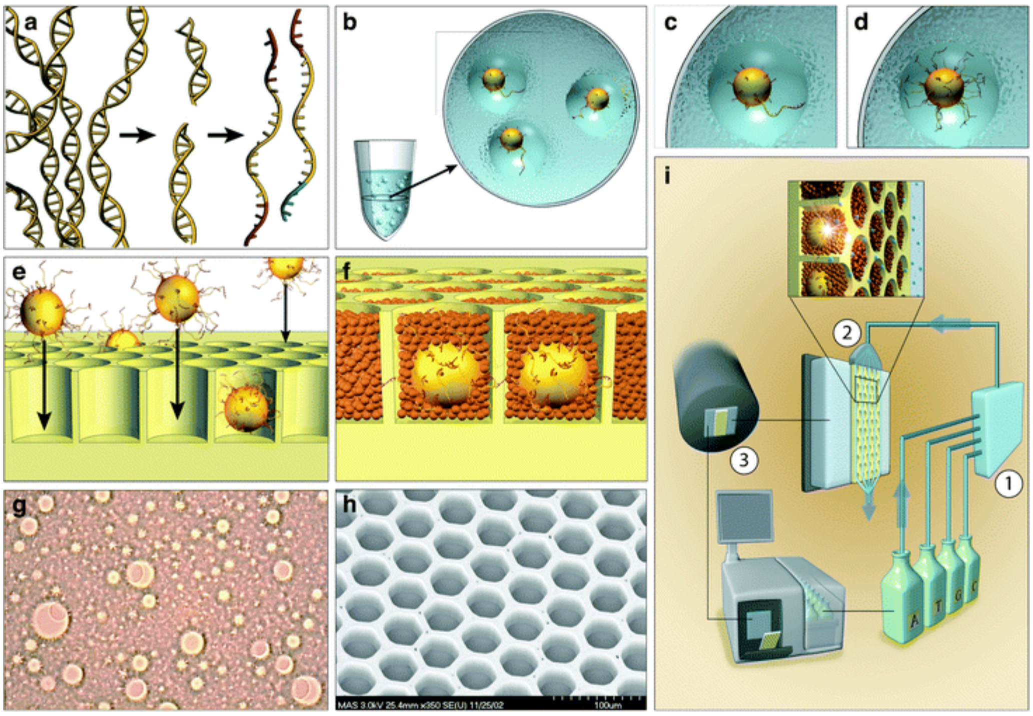


Figure 3. Pyrosequencing in 454 platforms. A) Genomic DNA is isolated, sheered and ligated to adaptors. B) Fragments are bound to beads, one sequence per bead. C) During emPCR, droplets capture single beads D) Amplification produces millions of copies of a unique DNA template. E) Beads are deposited into wells of a fibre-optic slide. F) Enzymes for pyrosequencing are added. G) Microscope photograph of emPCR. H) Electron micrograph of slide. I) Sequencing instrument consist of a flow chamber (1), a CCD sensor (2), and a CPU. Adapted from: Margulies *et al.*, 2005 (*34*).

The original *454 GS20* pyrosequencing platform achieved 100 fold the throughput of automated Sanger sequencing (*34*) but its next iteration, the *Roche 454 GS-FLX*, released in 2008 after Roche acquired the technology, allowed the sequencing of over 100 Mbp in 250 nucleotides (nt) long reads overnight. New reagents were released over the years, eventually reaching the 700 Mb mark with reads of 700-800 bp with the Titanium Plus upgrade. Nevertheless, 454 Life Sciences was shut down in 2013 by Roche after announcing it would discontinue the 454 platform by mid-2016 as it had become non-competitive (*35*).

During its relative short lifespan, the 454 sequencing platforms democratized access to metagenomic data as the arduous task of cloning was no longer a requirement and even larger collection of genomic pools from environmental samples could be swiftly and automatically processed. This was first demonstrated by Forest Rohwer's laboratory in 2006 in a study of the microbial ecology of water samples obtained from deep within an iron mine in Minnesota, USA (*36*). They managed to obtain over 70 Mbp of metagenomic data that were used successfully for 16S and functional profiling (WGS). They determined that for metagenomes about 1 in every 10^5 bases matched a 16S rRNA gene.

The massive profiling of 16S rDNA became especially popular as pyrosequencing allowed for cost-effectively surveying the taxonomy of environmental communities using PCR amplicons obtained with universal primers targeted to the hypervariable regions of the rRNA gene (*37*). This vastly expanded the understanding of the ecology in different niches from various geographical locations and multiple host organisms. Human-associated metagenomics was one of the most benefitted fields (*38*, *39*).

The growing interest in metagenomics also led to impressive massive metagenomic surveys such as the sampling of four oceanic regions by Rohwer in 2006 analysing 184 viral metagenomes (WGS) with pyrosequencing data (*40*) as well as Craig Venter's ocean expedition in which 6.3 Gbp were sequenced using cloning techniques (*41*). At the same time, new meta-omic fields started emerging, such as metaproteomics, first reported in a 2004 study by Wilmes and Bond (*42*) and metatranscriptomics in 2005 by Poretsky and collaborators (*43*). These new disciplines further expanded the techniques that could be used to study the environmental microbiota (Box 2) and further specialized into specific niches (*44*).

Box 2 – Meta-omics

The term "meta-omics", where the prefix *meta*- stands for "beyond" and the neologism -*omic* for a collective study, refers to the compendia of molecular and bionformatic techniques employed for characterizing specific layers of the microbiota. These include but are not limited to metagenomics, metatranscriptomics, metaproteomics and metabolomics (sometimes called "meta-metabolomics" but still with a *meta* approach), focusing on the study of genomes, transcripts, proteins and metabolites, respectively (*64*).

They are not restricted to single organisms or to culturable ones. Other meta-omics focus on specific groups of organisms and their nucleic acids such as the microbiome (in this case referring to living organisms, most commonly bacteria) (*79*), viromics (viral nucleic acids) (*378*), specific molecules as occurs with metamobilomics (mobile elements such as plasmids and prophages) (*379*), or the type of source material as ancient DNA in paleogenomics (*44*).

The decline of pyrosequencing came as collateral effect of the popularization of a different high-throughput sequencing technology also using a sequence-by-synthesis approach. Solexa's 1 G Genome Analysis System was the first commercial sequencer to reach the 1 Gb milestone in a single 72h run (*45*), although producing very short reads of ~ 30-50bp, which made it better suited for single genome resequencing (*46*). Illumina acquired Solexa in 2007 and produced the next iterations of the platform, eventually reaching up to 1 Tbp (tera [10^{12}] base pairs) with its Illumina HiSeq 2500 platform. It also commercialized read-length optimized platforms such as MiSeq that manages to produce over 10 Gb of ~500 bp reads. This specialization scheme gave Illumina an advantage in many fields, becoming the platform of choice for clinical trials, industry, research and human genome resequencing (recently reaching the $1000 genome milestone for cluster HiSeq X Ten) (*47*).

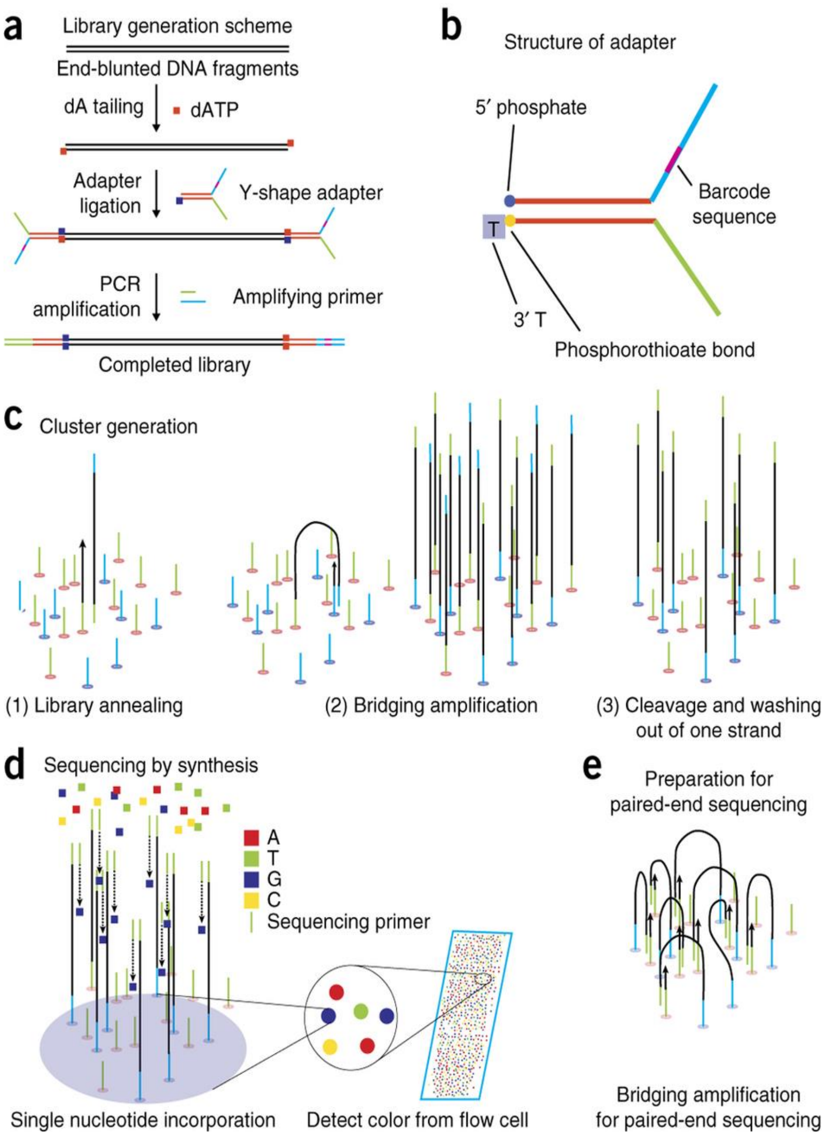

a
Library generation scheme
End-blunted DNA fragments
dA tailing
dATP
Adapter ligation
Y-shape adapter
PCR amplification
Amplifying primer
Completed library
b
Structure of adapter
5′ phosphate
Barcode sequence
T
3′ T
Phosphorothioate bond
c
Cluster generation
(1) Library annealing
(2) Bridging amplification
(3) Cleavage and washing out of one strand
d
Sequencing by synthesis
A
T
G
C
Sequencing primer
Single nucleotide incorporation
Detect color from flow cell
e
Preparation for paired-end sequencing
Bridging amplification for paired-end sequencing

Figure 4. Sequencing in Illumina platforms. A) Previously sheered DNA is flanked by adaptors. There are different types of library construction. In TrueSeq libraries the addition of dA prevents concatemerization before a Y-shaped adaptor is ligated. B) The adaptor is protected from nucleases and bears a barcode index sequence. C) Each sequence is attached to a flow cell by both ends using complementary adaptors and a complementary strand is synthesized. The 5'-3' strand is washed away and the bridge amplification process is repeated to create the clusters. D) sequencing primers are inserted matching the end of adaptors. Sequencing is done one nucleotide at a time using reversible terminators. A light emission is detected in each round. E) Optionally, paired-ends can sequence a second strand with another set of sequencing primers and the complementary strand. From: Shin *et al.*, 2014 (*48*).

Just like 454 technology, Illumina platforms depend on amplifying sequences to form clusters of each of the source sequences that can be then read by synthesizing the complementary strand (*49*). Original sequences are captured by 5’ and 3’ sequence adaptors that are covalently linked to a solid plate (Figure 4). They are then subjected to a few rounds of amplification that occur in a finite space, producing clusters each having the same molecule. Sequencing is then achieved by extending one nucleotide at a time for a fixed number of rounds. This works because each dNTP has a reversible terminator which is washed after each round to allow the sequencing to continue. Each time a nucleotide gets incorporated, it emits a photon that is captured by CCD sensor which identifies which nucleotide it is. More complex paired-end sequencing will be explained in the methods section but, briefly: two rounds of sequencing are carried out for sequencing longer reads, the first for sequencing the start of a longer DNA fragment, the second for sequencing the end of the complementary strand of the same sequence. Resulting pieces must be stitched together bioinformatically using the overlap between the two.

The large throughput of Illumina sequencers has helped metagenomic explorations move towards more ambitious metagenome (WGS) and metatranscriptome explorations as the price per base has become more accessible, currently < 100 € per Gb in a MiSeq and < 30 € in the latest HiSeq (*50*). The technology was at first used as an alternative or complementary approach in a classic work by Jeffrey Gordon’s lab in which they explored changes in diversity of the microbiota (454 sequencing) as a result of switching from low-fat to high fat diet, coupled with the analysis of the gene expression changes in the metagenome (Illumina sequencing) (*51*).

Over time, the increased length of the Illumina MiSeq (~500 bp inserts in 2x300 paired-end libraries) output was sufficient to render pyrosequencing’s ~700bp non-competitive. Not only the cost was just a tenth of the latter’s but it also produced almost 10 times more total bases per day (Illumina up to 15Gb per run; 454 ~ 0.7Gb per run). Most metagenomics and 16S profiling studies ultimately adopted Illumina as their main technology (*52*, *53*).

As of April, 2017, a third generation of sequencers is already available in the form of Oxford Nanopore Technologies’ *MinION* (a successor is on the works) and Pacific Biosciences *RSII* and *Sequel* platforms (*54*, *55*). They all use a single-molecule real-time sequencing approach that is theoretically unbiased as it does not require previous amplification.

The *MinION* is marketed as a handheld USB-powered inexpensive apparatus which was created with field experiments in mind. It measures ionic current changes produced as DNA passes through an array of protein nanopores to determine the sequence and it can theoretically process continuous reads of up to >100Kb but latest reports point out the median length in fact stands around 1Kb with ~38% error rate (*54*). In contrast, PacBio’s sequencers use immobilized DNA polymerases in microwells, each processing a single strand of DNA by adding fluorescent-labelled dNTPs that can be monitored as the sequencing progresses (*55*). The usage of the RS II system with P4/C2 reagents reported median length of over 1400 bp and higher accuracy than that of 454 experiments when a consensus is predicted (*56*).

Both third generation approaches have yet to be fully tested for metagenomics. So far, there are only reports of PacBio’s *Sequel* platform being used in experiments sequencing the complete 16S rRNA gene sequences which, contrary to the Illumina regular experiments, achieves species-level resolution in bacterial taxonomic identification (*57*). The MinION sequencer has been used for in-site trials in Antarctica to amplify genomic DNA (gDNA) from environmental samples with less than optimal results but was more successful in the sequencing of Ebola viruses in Guinea (*58*). The technology seems promising but still requires much work.

Studying microbiota

Resident microbiota

Any biological niche, be it environmental or host-associated, is inhabited by microbes spanning all three domains in the tree of life (Bacteria, Archaea, and Eukarya commonly addressed as the microbiota as a whole) as well as viruses (*59*). The varying composition and interactions between them and their environment results in divergences in their genomic, transcriptomic, proteomic, and metabolic composition that drive their adaptation (*60*). Moreover, a great part of this diversity is non-culturable (Figure 5).

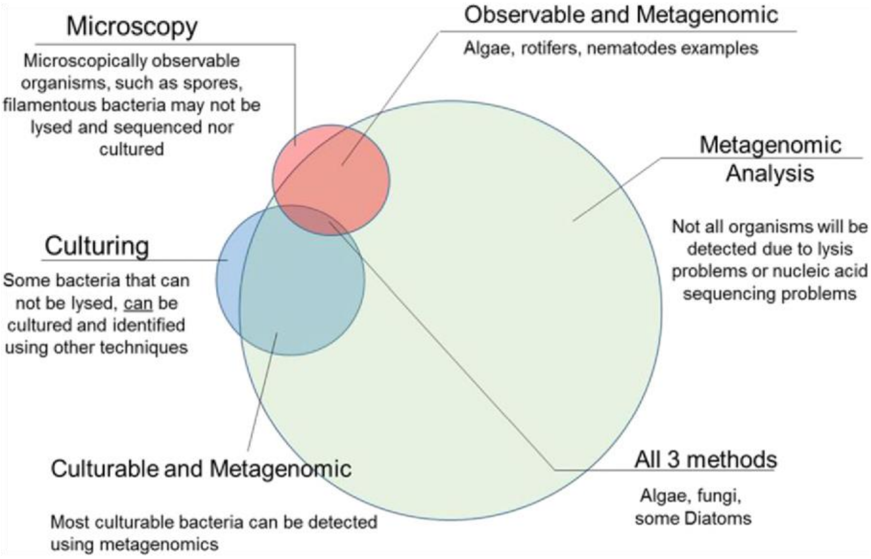

Microscopy
Microscopically observable organisms, such as spores, filamentous bacteria may not be lysed and sequenced nor cultured
Observable and Metagenomic
Algae, rotifers, nematodes examples
Metagenomic Analysis
Not all organisms will be detected due to lysis problems or nucleic acid sequencing problems
Culturing
Some bacteria that can not be lysed, can be cultured and identified using other techniques
Culturable and Metagenomic
Most culturable bacteria can be detected using metagenomics
All 3 methods
Algae, fungi, some Diatoms

Figure 5. Relationship among different methods of microbiota detection. Traditional methods of microscopy and culturing fail to detect a large proportion of the microbial population in a sample. Metagenomic approaches miss bacteria that cannot be lysed with simpler protocols or are spores. From: Tighe *et al*., (*61*).

Resident microbiota is, thus, far from static as it continuously adapts to its biome by specializing, adjusting to available types of carbon resources while overcoming internal struggles (*62*). Within any microhabitat, competing organisms thrive and die in regular cycles or as result of external stimuli (*63*). They unleash their microbial warfare or form complex relations, synergically complementing metabolic pathways and they ultimately provide these intricate ecosystems with nutrients and enzymatic products (*64*).

All these interactions occur naturally within the boundaries of microscopic habitats that over the course of about a decade have captivated the attention of the scientific community. The advent of metagenomic studies opened the possibility of identifying and studying the important players, their roles and contributions in multiple scenarios but it was only the beginning. The different techniques of multi-omic approaches can only provide partial pictures of a moving and very active world that calls for further exploration that can only be achieved by integrating multi-layered data on species, proteins and molecules, and the interaction between them and their habitat (*65*). This is especially important in host-associated niches as some of these may cause a significant impact in overall health of the host (*66*).

Human microbiota

It should not come as a surprise that the microbiota that resides within the human body has been the most comprehensively studied, given its economic and academic interest. The notion of it forming a holobiont with its host (Box 1) is very relevant to homeostasis, and has established a new conceptual framework to study the reciprocal role of our inner microbial world which has been suitably addressed as a "human forgotten organ" (*67*, *68*), with its own unique set of genes and other genomic features that altogether comprise "our *other* genome", the human metagenome (*69*). In a humbling note, this enormous genetic collection far exceeds our own as it bears more than a hundred times the number of protein-coding genes found in the 3.2 Mbp human genome (*70*) which in turn has ~19,000 of them (*71*). In a little more than a decade, hundreds of finished and ongoing research projects have tackled the arduous task of untangling the complex relationships that exist between resident microorganisms and their human hosts (*16*, *66*, *72*, *73*). A specifically important effort has been made towards defining the composition of "healthy" microbiota states (*74*) and their deviations from these states known as "dysbiosis" (*75*).

The different microorganisms that reside within human niches can interact with their host in multiple ways. Although most of the initial microbiome projects focused on pathogenic bacteria (either strict or opportunistic, that negatively affect host health when they thrive), it soon became clear that the majority of the human body's residents lived in symbiosis, a relation between the species and host that granted at least one of them benefits without harming the other (either mutualistic or not), or commensals, organisms that coexists without detriment but providing no benefit either (*76*).

In his 1977 review of the microbial ecology of the gastrointestinal tract, Dwayne Savage suggested there may be about 10% human cells (10^{13}) in a normal human organism, with the rest belonging to prokaryotes (10^{14}) (*77*), most of them residing in the gut. Although this pioneering calculation was only meant to make a point back then, the concept of the human body having more bacteria than human cells has transcended, inexorably reaching out to the general public, in part thanks to the rising popularity of human metagenomic studies in the past decade. More importantly, it has raised awareness on the importance of microbiota and has started a constant, though slow, shift from the concept of the human as a host to that of a holobiont harbouring whole complex microenvironments.

The 1:10 ratio of human to bacterial cells has been recently revised by Sender and collaborators, suggesting the actual figures may be closer to a 1:1 proportion, with mean estimates of $3.8 \cdot 10^{13}$ bacteria and $3.0 \cdot 10^{13}$ human cells in a reference male (20-30 years of age;70 kg;170 cm) or roughly 10:1 if only nucleated cells are considered in human (*78*) as shown in Figure 6. Additionally, they provided updated estimates on the total biomass, with bacteria accounting for 0.2kg, a mere 0.3% of the reference's total weight. Regardless, numbers are approximate and never as biologically relevant as the underlying diversity and the role that resident microbiota may play.

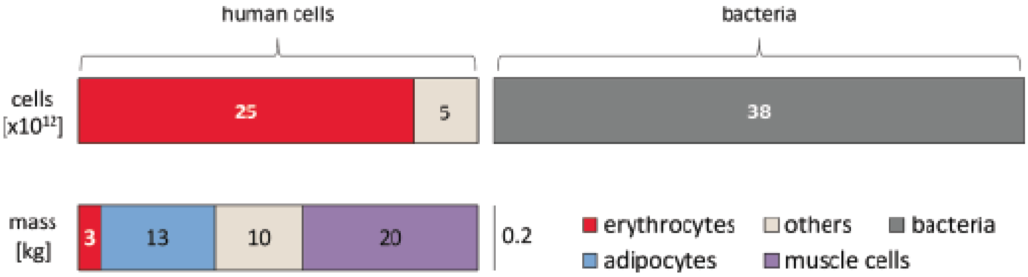


Figure 6. Distribution of cell number and mass for different cell types in the human body (70 kg adult man). The bar on top displays cell totals. Erythrocytes comprise 84% if all human cells while muscle cells represent just 0.0001% and adipocytes 0.2%. The bar below compares total weight. From: Sender *et al.*, 2016 (*78*).

Unsurprisingly, a large proportion of the human-associated metagenomic studies have focused on the highly-populated human gut (samples from duodenum/jejunum/colon/faeces), although many other have explored the oral cavity (dental plaque/saliva), skin, vagina, respiratory tract, stomach, ear, eye, among several others (*79*). As a result, bacterial and archaeal diversity within the human body has been thoroughly mapped, establishing reference "core genes" and "core microbiota", basal sets of genes or conserved functions and their taxonomic clusters specific to a particular niche (*51*, *69*). With all that, studying the human gut microbiome is key in order to understand the interactions of the microbiota in larger scale and has been the main focus of large-scale international collaborations (see Box 3).

Box 3 - Large-Scale Projects

The four oceanic regions by Forest Rohwer and collaborators (*40*) was the first large-scale project, producing a large set of new bacteriophages sequences (180Mbp). A year later, the Global Oceanic Sampling expedition by Craig Venter (focused on bacteria) obtaining over 6Gbp, enabling partial genome reconstruction of some species (*41*).

The Metagenomics of the Human Intestinal Tract (MetaHIT; 2008-2011) was a 22 million € initiative founded by the European Union where eight countries collaborated towards unveiling the composition of the human gut metagenome using a high-throughput WGS approach, producing nearly 600Gbp (*69*). Their two major findings were the establishment of a basic catalogue of metagenomic genes, consisting of 3.3 million items which was the result of the largest metagenomic survey until then, and the proposal of three main enterotypes, general microbiota configurations characterized for their predominant bacteria (*Bacterioides*, *Prevotella* or *Ruminococcus*) (*80*).

Complementary to this study was the Human Microbiome Project (HMP; 2007-2011), a 115 million dollar 5-year project founded by the American National Health Institute (Figure 7), aimed at exploring the diversity and dynamics of the metagenomic component of multiple human niches (the most important being the gut, oral cavity, nose, skin and vagina) (*81*). While the MetaHIT project focused on WGS, this one was heavily oriented towards 16S profiling and included the sequencing and assembling of new reference genomes. Their aim was to provide and standardize a robust framework for processing metagenomic samples and provide reference pictures of healthy microbial diversity and the way it is altered over the span of approximately 2 years. It produced over 3.5 Tbp of DNA sequences (*82*).

The Integrated Human Microbiome Project (iHMP; 2013-2016) was announced as a four-year second phase of the HMP, with a funding of 7.5 million dollars from the NIH. Three vertices composed this initiative, focusing on the dynamics of preterm birth, inflammatory bowel disease, and type 2 diabetes, all of them integrating data from multiple meta-omics and complementary techniques including 16S profiling, metagenomics, metatranscriptomics, lipidomics, cytokine assays, interactomics, metabolomics, viromics and single-cell analysis, among others (*83*). The results of this project are still pending major publications.

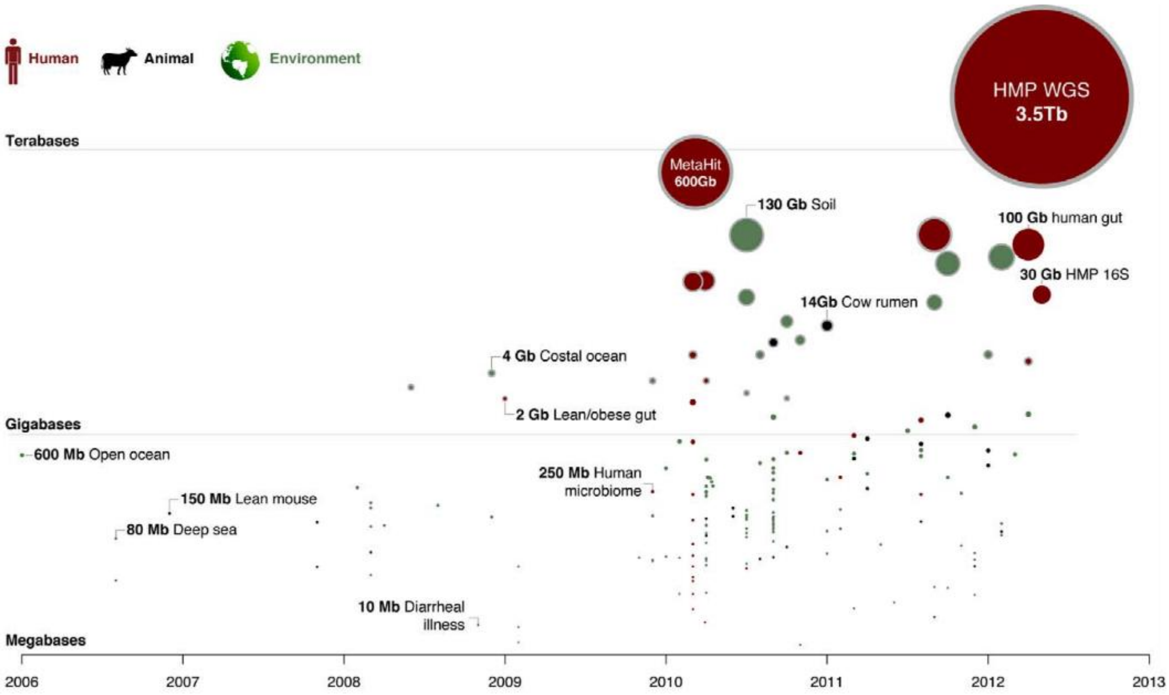


Figure 7 Timeline of microbial community studies using high-throughput sequencing up to 2012. Each circle represents a 16S or WGS bioproject in NCBI (May 2012) indicating the amount of sequence data produced at the time of publication. Projects are colour-coded and selected projects are labelled. From: Gevers *et al*., (*84*).

The Integrated Human Microbiome Project (iHMP; 2013-2016) was announced as a four-year second phase of the HMP, with a funding of 7.5 million dollars from the NIH. Three vertices composed this initiative, focusing on the dynamics of preterm birth, inflammatory bowel disease, and type 2 diabetes, all of them integrating data from multiple meta-omics and complementary techniques including 16S profiling, metagenomics, metatranscriptomics, lipidomics, cytokine assays, interactomics, metabolomics, viromics and single-cell analysis, among others (*83*). The results of this project are still pending major publications.

The Earth Metagenome Project (EMP; 2010-ongoing) is an international collaboration aiming at studying microbial life on all types of niches in our planet, over ~200,000 microbial samples and over 500,000 genomes by means of sharing results of ongoing individual projects (*85*). Most of them are 16S profiling projects (*86*) from more than 40 different biomes including human-associated samples and environmental samples related to human activity. As of April 2017, more than 130,000 samples were available online from over 300 studies (*87*).

The Extreme Microbiome Project (XMP; 2014-ongoing) is an initiative from the ABRF group to study extreme and unique environments in multiple locations around the world (*61*). This collaboration aims at identifying genetic adaptations of bacteria and archaea living in hostile environments and will greatly increase the number of archaeal reference sequences.

Knight Lab’s American Gut Project (started in 2013-ongoing) is an initiative with a crowd-funded model asking for public participation for sample gathering (*88*). These ultimately working as a private microbiome sequencing service. Results have helped researchers study the association of microbiota diet, ethnic origin, and geographical location. The project has spawned spinoffs in the United Kingdom, Australia and Asia.

The Metagenomics and Metadesign of the Subways and Urban Biomes (MetaSUB; 2016-2020) is an urban-sampling initiative to study microbiota in human habitats using 16S and WGS, mainly in transportation systems hubs of 51 cities from 31 countries (*89*). This project is based on the success of PathoMAP (2013-2015), the first City-Scale Metagenomics effort, which analysed New York (*90*).

Gut to Thrive or Die

In the boom of meta-omics, the gut microbiome has seen the greatest number of studies and has concentrated the largest number of important discoveries in host-associated microbiota. In 2010, the MetaHIT (see Box 3) consortium published a cornerstone article with the results of a large-scale project to survey stool samples (which are commonly used for studying the colon microbiota) of mixed population of 124 healthy and diseased European individuals (576.7 Gb of sequence) (*69*). Their findings supported the existence of over 3.3 million nonredundant genes in the complete set, ~150 times more than its human counterpart, 99% of which presumably belonged to bacteria. In spite of it not being an exhaustive catalogue, they observed a core set of species common to most individuals but found a larger genetic variability.

The Firmicutes and Bacteroidetes phyla were dominant (over 90%). Also prevalent in their analysis were *Dorea /Eubacterium/ Ruminococcus* groups as well as bifidobacteria, proteobacteria and streptococci/lactobacilli groups whilst basic functions were housekeeping and those required for niche survival. A second functional group was reported to contain gut-specific functions that contributed to the ecosystem: adhesion to the host and harvesting of sugars, including degradation and uptake pathways for pectin and sorbitol (found in fiber-rich diets but not directly usable by humans) and fermentation of mannose, fructose, cellulose and sucrose, in part to short-chain fatty acids (SCFA), that can be used as energy source by the host, as well as amino acid and vitamin (biotin, phylloquinone) production (*69*). Additionally, gut bacteria degrade xenobiotics, including non-modified and halogenated aromatic compounds. Importantly, the MetaHIT consortium also reported the prevalence of prophage related sequences (~5%) in the metagenome, suggesting bacteriophages play an important ecological role in microbial dynamics.

The metabolism of SCFAs has since become a key element in functional analyses of the human-associated microbiota. These are composed of one to six carbon aliphatic organic acids, end products of the dietary fibre fermentation. The most important are the acetate, propionate and butyrate, readily present in the colon in an approximate molar ratio of 60:20:20 (*91*).

In a second landmark bioinformatic study from the MetaHIT consortium, Arumugam and collaborators explored the functional diversity of the human gut microbiota by integrating 39 newly sequenced (mostly Sanger) into a pool of available host-associated metagenomes (*80*) to study clusters of orthologous groups (COGs) composition. A large proportion (40-60%) of the metagenomic sets did not map to any COG. Although most functions mapped dominant species, the contribution of low-abundance (minority) groups clearly suggested abundant species or genera cannot reveal the entire functional complexity of the gut microbiota alone.

The authors also reported the detection of three species-driven groupings or enterotypes in their gut microbiota datasets, each distinguished by levels of *Bacterioides*, *Prevotella*, and *Ruminococcus*, respectively, though the third group did not hold perfectly when external data were used. The *Bacteroides* enterotype was enriched in enzymes related to carbohydrate and protein fermentation. The *Prevotella* enterotype also presented *Desulfovibrio*, and was thought to be focused on the degradation of mucin glycoproteins present in the mucosal layer. The *Ruminococcus* enterotype, also rich in *Akkermansia*, had mucin degraders and was enriched in membrane transporters to import sugars. Finally, the largest differences in the MetaHIT datasets were found to be age and nationality dependent.

In a subsequent cross-sectional study with controlled diet, Wu and collaborators reported the *Bacteroides* and *Prevotella* enterotypes as strongly associated with long-term diets rich in protein and animal fat, and fibre, respectively (*92*). A study of European and African children obtained similar results by indirectly comparing western diet (protein and fat rich) with rural diet (fibre based and low in protein) (*93*), therefore suggesting the gut microbiota seems to be very susceptible to carnivorous (promoting a *Bacteroides* enterotype) or vegetarian diets (promoting a *Prevotella* enterotype).

In spite of its general acceptance, the existence of enterotypes has been firmly questioned over the years (Figure 8). As more data emerge, evidence accumulates supporting that there is a continuous gradient of dominant taxa rather than the discrete enterotypes that were originally proposed (*94*). As Knights and collaborators pointed out in their report, enterotypes effectively mask meaningful variation occurring within clusters magnified where there is lack of sampling between the extremes. Furthermore, they demonstrated with longitudinal data that a subject’s microbiota may traverse different enterotypes at varying timepoints. Nonetheless, enterotypes are still a main subject of research in contemporary studies (*95*, *96*).

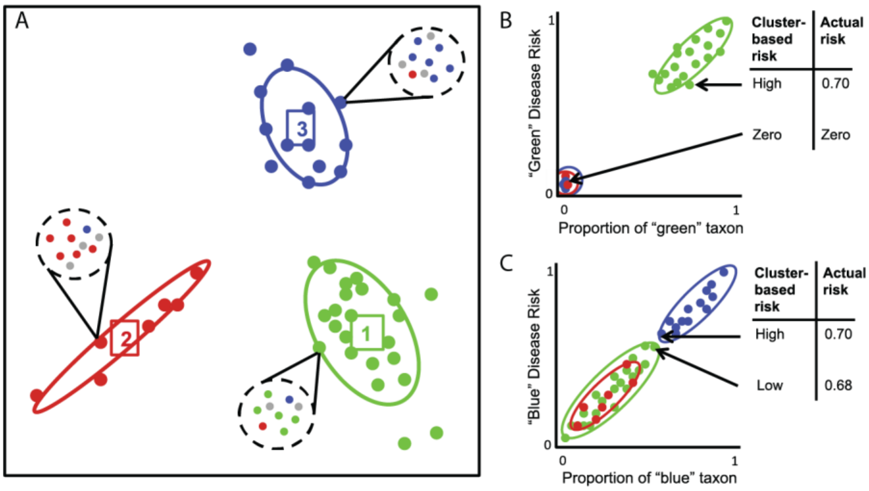


Figure 8 Clustering continuous data may mask within-cluster variation. A) Hypothetical clustering of complex bacterial communities. Green is healthy condition, blue and green represent diseases. Most taxa are present in every cluster, except for the green taxon which only appears in cluster one. B and C are plotted on a continuous axis showing proportion of a given taxon. B) When disease risk is correlated with taxa found in only one cluster (green), difference between enterotype are magnified and variation within the disease cluster is masked. C) When disease risk is correlated with taxa found in more than one cluster (blue), clustering masks important risk variation within clusters. From: Knights *et al*., 2004 (*94*).

The complete picture

In 2012, the HMP consortium (see Box 3) published the aggregate results of 5 years of metagenomic studies achieved by high-throughput sequencing (both Illumina and 454 platforms were used) over 800 (goal was 3000) new reference strain genomes of organisms (mostly bacterial but also including viral and eukaryotic genomes) and the sequencing of over 5,000 16S profiles from 242 healthy adults surveying 15 to 18 body sites at three different timepoints over 22 months. Some selected individuals in the varying niches were sampled for 680 WGS sets from a subset (*84*). It was the most exhaustive survey up to that point.

In an effort pinpoint the most important contributors to human health in the microbiota, the HMP consortium found that the diversity and abundance of each of the surveyed niches vary widely even among healthy subjects (Figure 9), having strong niche specialization both within and among individuals as no taxa was universally present in all body habitats (*97*). Even though the microbiota composition varied among individuals, in part due to diet, environment and host genetics, metabolic pathways were mostly conserved. Complexity also varied among sampled sites, with vaginal microbiota having a simpler composition (~16,000 protein families), whilst faeces and the oral cavity displayed the largest (~400,000 families).

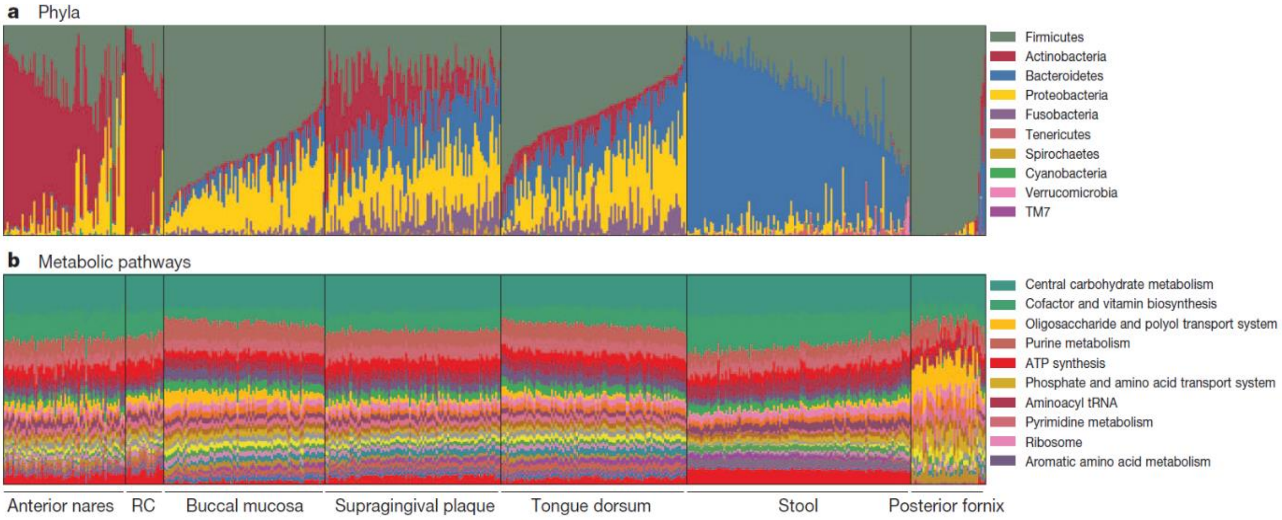


Figure 9. Taxonomic and functional composition. Vertical bars represent microbiome samples by body habitat in the seven general locations having both 16S profiles and WGS data. RC, retroauricular crease A) Abundance of different phyla b) Most abundant pathways. From: Huttenhower *et al.,* 2012 (*97*).

Even though no single "healthy" configuration had been found, composition and abundance of the most prevalent organisms in a host was reported to fluctuate in an ultimately predictable pattern as the core set of organisms in each niche seemed to be stable over time (*97*), contrary to the transitory and highly variable minority fraction. The HMP consortium also reported the absence of priority pathogens from healthy individuals (class A-C pathogens defined by the American National Institute (*98*)) while opportunistic pathogens (defined by the Pathosystems Resource Integration Center (*99*)) were nearly ubiquitous (>1% prevalence of >0.1% abundance).

In regard to metabolic pathways in the HMP datasets, these were far more conserved than their taxonomic counterpart, with the most important (core pathways) being for ribosome and translational machinery, nucleotide charging and ATP synthesis, and glycolysis (*97*). The oral microbiome was the most variable, showing enrichment in transport of phosphates, mono- and disaccharide, and amino acid transport in the mucosa, as well as synthesis of lipopolysaccharides and spermidine/putrescine on the tongue and dental plaque, supporting the hypothesis of niche specialization (Figure 10). Association with host phenotypic traits was reported for ethnicity and bacterial distribution in most niches. Also, age correlated with pathways in skin and low pH with lactobacilli in vagina but even the strongest had low effect sizes and a large proportion of unexplained variance.

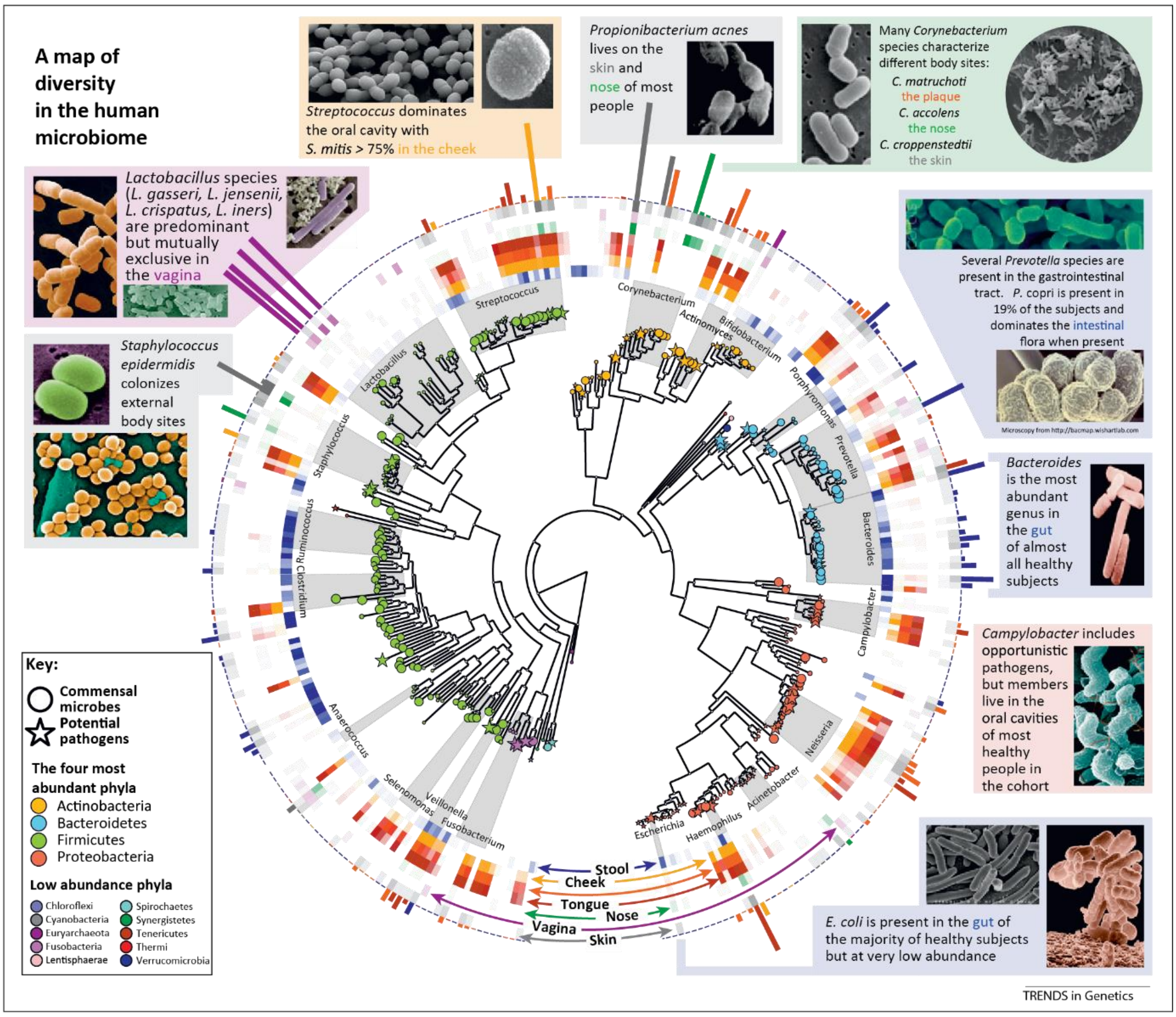

A map of diversity in the human microbiome
Streptococcus dominates the oral cavity with S. mitis > 75% in the cheek
Propionibacterium acnes lives on the skin and nose of most people
Many Corynebacterium species characterize different body sites:
C. matruchoti the plaque
C. accolens the nose
C. croppenstedtii the skin
Lactobacillus species (L. gasseri, L. jensenii, L. crispatus, L. iners) are predominant but mutually exclusive in the vagina
Several Prevotella species are present in the gastrointestinal tract. P. copri is present in 19% of the subjects and dominates the intestinal flora when present
Microscopy from http://bacmap.wishartlab.com
Staphylococcus epidermidis colonizes external body sites
Bacteroides is the most abundant genus in the gut of almost all healthy subjects
Campylobacter includes opportunistic pathogens, but members live in the oral cavities of most healthy people in the cohort
E. coli is present in the gut of the majority of healthy subjects but at very low abundance
Streptococcus
Corynebacterium
Actinomyces
Bifidobacterium
Porphyromonas
Prevotella
Bacteroides
Campylobacter
Neisseria
Acinetobacter
Haemophilus
Escherichia
Fusobacterium
Veillonella
Selenomonas
Anaerococcus
Clostridium
Ruminococcus
Staphylococcus
Lactobacillus
Stool
Cheek
Tongue
Nose
Vagina
Skin
Key:
Commensal microbes
Potential pathogens
The four most abundant phyla
Actinobacteria
Bacteroidetes
Firmicutes
Proteobacteria
Low abundance phyla
Chloroflexi
Cyanobacteria
Euryarchaeota
Fusobacteria
Lentisphaerae
Spirochaetes
Synergistetes
Tenericutes
Thermi
Verrucomicrobia
TRENDS in Genetics

Figure 10 Map of diversity in the human microbiome. The human microbiome is dominated by four phyla: Actinobacteria, Bacteroidetes, Firmicutes and Proteobacteria. A taxonomic tree of the most abundant organisms is in the centre. Commensal microbes are indicated by circles, and pathogens are indicated by stars. The middle ring marks the niches where the organisms were found. Bar heights on the outside circle are proportional to taxa abundance at the body site of greatest prevalence. Intensity of external colours reflects prevalence. From: Morgan *et al.*, 2013 (*100*).

The overall stability of the system tends to be conserved over long periods of time (*97*) but its abundance and composition reacts to external stimuli such as the introduction of foreign species and physiological changes of the host, and it is capable of surviving external threats such as viruses and, occasionally, the own host's immune system and feeding conditions (*101*, *102*). After all, inner relations within the microbiota are far more complex than just binary variables reflecting pairwise associations, or else presence/absence of particular taxa or functions. There is a fierce competition for nutrients and synergetic survival efforts, cyclic or procedural episodes of colonization and biofilm formation, overall resulting in very different organisms thriving and dominating the landscape at any given moment (*94*). The general view is that cooperation promotes overall system efficiency in community-based ecological networks (*70*) by syntrophic interactions (cross-feeding, where one species feeds on the product of other). However, this does not necessarily imply there may be a beneficial effect on secreting species. In fact, Foster and Bell did not find positive gains when different species were put together in mixed cultures. They actually reported net negative effects supporting a competitive interaction may be dominant (*103*).

This may be explained by a more recent interpretation, called the Black Queen hypothesis, appealing to reductive genomic evolution driven by genetic drift (*104*). This theory assumes some gene functions are dispensable (which is supported by the detected functional redundancy (*97*) so that gene loss can provide a selective advantage when these functions can be provided by neighbouring species. As selective pressure relaxes and genes are lost for a given organism, it also leads its eventual metabolic interdependence by increasing reliance on molecules that are no longer produced by specific microbes at the same time it promotes waste as well as exploitation by opportunistic commensals (*105*). Cooperation is still expected to be dominant when the genetic pool is not as extensive in the community's metagenome.

A recent ecological study from Coyte and collaborators reported that high diversity of species may coexist in stable conditions when the system is dominated by competitive interactions, contrasting to cooperating systems (*101*). This is explained because, although competition may decrease the systems overall fit, it reduces the destabilizing effect of cooperation given by codependence of its community members (when the abundance of a connected member in a community network is reduced, it usually pulls down others with it).

Nevertheless, the actual presence of some of these microorganisms is necessary for their host. Their contribution towards a viable protein and metabolic profile ranges from fermentation and degradation of nutrients that cannot be processed by the host alone (*69*) to the modulation of the immune response as the host acts over a system rather than on individual strains or species due to the functional redundancy (*106*). Some studies are even interested in the potential behavioural tuning of the host's behaviour (gut-brain axis), presumably produced by the microbiota in a bidirectional communication with cognitive centres through immunological and neuro-endocrine systems associated with stress response, anxiety and memory (*107*).

Establishment and maturation of human microbiota

Mammals had been historically thought to be born sterile, with neonates having their first exposure to microbes during delivery (*108*). Such a pivotal moment in the development of the immune system has been the subject of study of many groups, setting to understand the onset of specific species of microbes and their maturation towards more stable configurations.

By the turn of the millennium, this view had already been challenged by the detection of bacteria in umbilical cord of healthy neonates and murine amniotic fluid but it was a classical experiment by the group of Jiménez and collaborators that proved the first colonization may occur differently (*109*). They studied the meconium (first deposition of a newborn, which is not formed by ingested nourishment and is formed during foetal development) of mice pups aseptically extracted prior to labour. These bore the same marked *Enterococcus faecium* strain (previously isolated from breast milk of a healthy woman) that were fed to the pregnant mice, suggesting microbiota can pass the placental and amniotic barriers to reach the foetus and may be vertically inherited.

An extensive study of the microbial contents in the human placenta by Aagaard and collaborators demonstrated the composition (rich in Actinobacteria, Firmicutes, Fusobacteria and Proteobacteria) was more similar to that from the oral cavity than to vaginal microbiota (*110*), obtaining similar results to some experiments that had been previously carried out in murine models but considered to be indicators of intrauterine infections (*111*). This may suggest that the placental microbiome may be established by the haematogenous (blood formation) spread, and delivered during early vascularisation and placentation.

The method of delivery provides differential colonization of microbes in the first weeks of life (*112*). Infants born via vaginal delivery are exposed to microbiota from the birth canal and the gut, establishing a basal microbiota configuration that initially resembles that of the mother's vagina, bearing elevated numbers of *Lactobacillus*, *Prevotella*, and *Sneathia* spp. whereas microbiota from infants born through caesarean section resembled that of the mother's skin with *Staphylococcus*, *Corynebacterium*, and *Propionibacterium* spp. (*113*). Furthermore, caesarean born infants had been reported to show reduced diversity during the first two years of life and colonization by Bacteroidetes may be delayed for over a year and the lack of bacteria transmitted through vaginal delivery was hypothesised to favour enterococci colonisation and the maturation of the immune response (Th1-like responses) (*114*). Breast milk is also a source of microbes for infants. Although it was first considered to be sterile, it is now known to be rich in *Staphylococcus*, *Pseudomonas*, *Streptococcus* and *Lactobacillus* and it has recently been proposed that it may be a fail-safe mechanism for mothers to pass along their bacterial imprint (*115*). A recent longitudinal study by Chu and collaborators demonstrated that all body sites in the neonate presents similar composition of microbiota until six weeks after delivery, after which there was no observable difference in the functional profile of the neonate microbiota regardless of delivery mode and there was discernible niche specialization (*116*).

Perhaps, the major role the mothers' early colonizers is related to the maturation of the immune response. The innate immune system recognizes bacteria using pattern recognition receptors (PRRs) that detect microbe associated molecular patterns (previously thought to be pathogen-exclusive) and the composition of early microbiota may train cells into tolerating certain microbes, avoiding inflammatory responses when regular commensal species are found (*117*).

Microbiota also varies with age (Figure 11). Infant microbiota composition is highly volatile and does not stabilize until after at least three years, as Yatsunenko and collaborators proved using large cohorts of babies from various parts of the world. The most notorious change in early life microbiota results from solid food introduction to the diet as well as antibiotic intake (*118*). On the other extreme, aging microbiota has been characterized in a large cohort of 178 elderly people (*119*), resulting in the detection of a higher inter-individual variability of the composition when compared to young adults and correlation to reduced health markers.

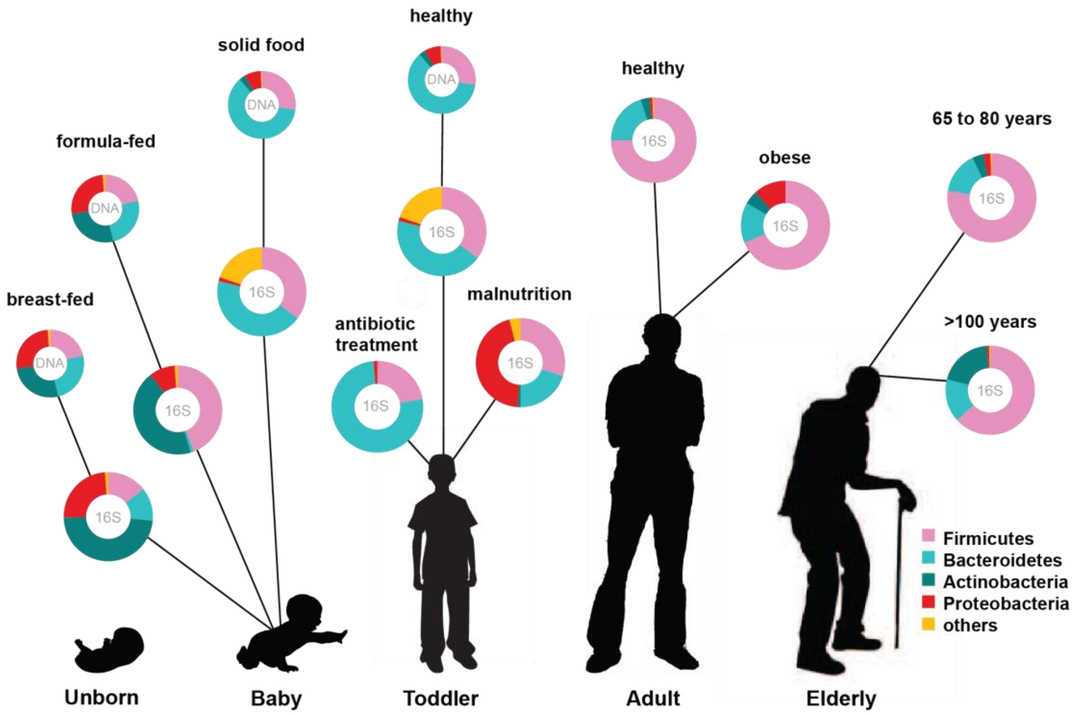


Figure 11. Onset and shaping of microbiota through life stages and perturbations. An overview of the abundance of key phyla is shown at different stages in life. Data from different studies using 16S profiling and WGS was employed. From: Ottman *et al.*, 2012 (*120*).

Oral fixation

With regard to microbiota in human-associated niches, the oral cavity has been one of the most studied by the scientific community in the past decade, second only to the populations inhabiting the gut (*82*). Whereas the skin has a mean surface of 1.8m^2 that is permanently exposed to external organisms (*121*), the oral cavity is the main entry for most of them into the gastrointestinal tract and, in general, to the human body. Having a mean surface of 214 cm^2, a volume of ~159 cm^3, and 0.77-1.07mL of saliva (*122*, *123*), the human mouth holds very distinctive communities and poses different ecological challenges to the microbes that inhabit it (*124*).

Anatomy and Physiology of the Human Oral Cavity

To fully understand oral microbiota, it is important to describe its niche, starting with the anatomy and physiology of the oral cavity. The following descriptions have been summarized from *Gray's Anatomy: The Anatomical Basis of Clinical Practice* (*125*):

The mouth, or oral cavity, is an orifice lodged between the maxillae, the largest pair of pneumatized (porous) bones in the midface and the mandible, both bearing the teeth in the alveolar processes (which are socketed bone ridges where teeth are found). Internally, the mouth extends from the lips and cheeks to the tonsils and the anterior pillars of the fauces, connecting the oropharynx (Figure 12). The teeth separate the external vestibule from the internal oral cavity proper while the roof of the mouth, the palate (anterior hard palate and posterior soft palate), separates the oral and nasal cavities. The floor, occupied mostly by the tongue, is formed by muscular tissue. Lateral walls are limited by the cheeks and retromolar regions. Three major pairs and numerous other minor salivary glands are distributed in the oral cavity. Muscles are associated with the lips, cheeks, floor of the mouth and tongue.

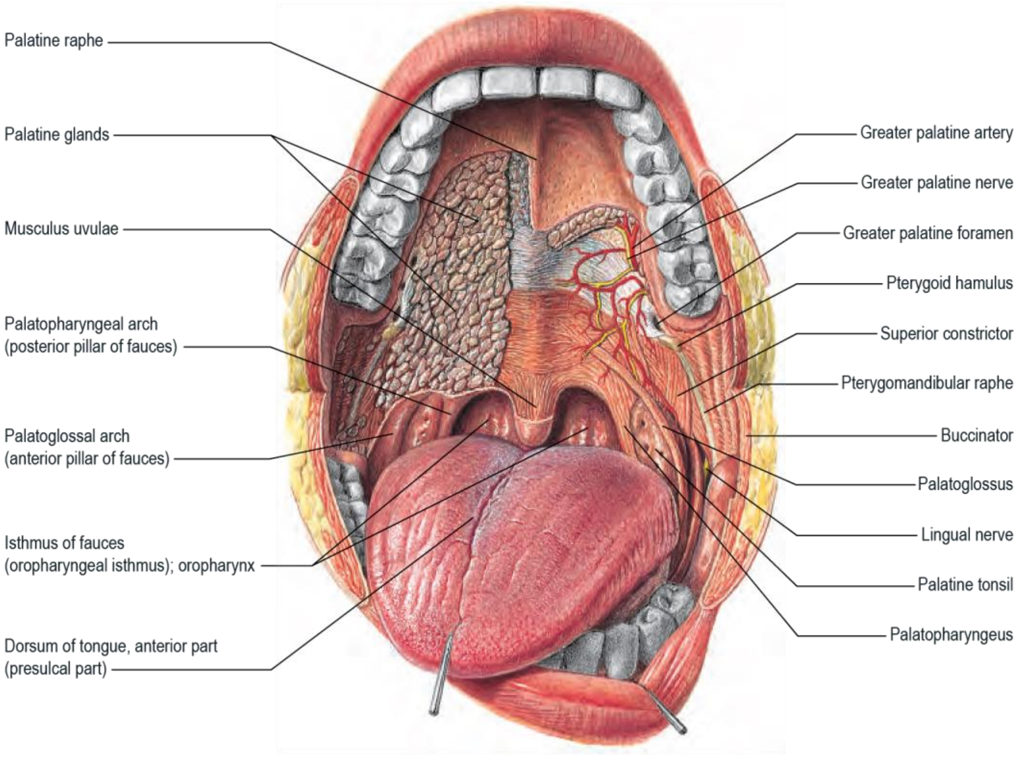


Figure 12. The oral cavity, oropharyngeal isthmus and muscles of the palate. From: Gray's Anatomy 41th Ed (*125*).

Most of the oral cavity is covered by mucosa that protrudes from the skin in the labial margins. The inner mucosa in the cheek is adhered to the buccinator muscle allowing stretching with movement and may present visible sebaceous glands. The labial mucosa has elevations caused by underlying mucous glands. The oral vestibule runs in the space between teeth and cheeks/lips, covering the alveolus of the jaw in a through called fornix vestibuli, traversed by connective tissue folds forming arches, the most prominent being the labia frena (upper and lower frenula) and those near the canines or premolars.

The mucosa in the oral cavity can be sub-classified into three different types:

The lining mucosa is formed by stratified squamous epithelium, presenting fat deposits and mucous salivary glands. It has red colour and covers the ventral surface of the tongue, floor of the mouth, alveolar processes, and the internal surfaces of cheeks and lips. The alveolar mucosa is loosely attached to the alveolar bone and is darker in colour due to underlying blood vessels. Finally, the paler gingival mucosa covers the upper part of the alveolar bone and the teeth cervical region (tooth neck).

A second type, the masticatory mucosa, is formed by the gingivae, palate and tongue dorsum. It is affected by the physical friction and pressure producing keratinization of the mucosa that is characterized by the hardening of the outer epithelial layer and general whitening of the area. Keratinization occurs as a result of cells undergoing keratinocyte differentiation, maturing and programmatically dying to convert themselves into tough cornified squames which may have nuclei (parakeratinization) or not (orthokeratinization). The gingivae (or gums) are differentially attached to the periosteum (outer membrane covering a bone). The attached gingiva is stippled (texturized due to normal connective tissue projections) whereas the free gingiva, having a looser grip of the bone, is the one mm area around the teeth and is non-stippled (smooth). The area between adjacent teeth is called interdental papilla.

The tongue is divided by a fibrous septum attached to the hyoid bone. Four pairs of extrinsic muscles (genioglossus, hyoglossus, styloglossus and palatoglossus) control its movement while intrinsic ones (superior/inferior longitudinal, transverse and vertical) alter its shape. The tongue is traversed by a v-shaped groove called the sulcus terminalis (terminal sulcus), found two thirds to the back (Figure 13). The anterior part, the presulcal dorsum, is covered by a third type, specialized mucosa, containing four kinds of epithelial projections or lingual papillae. Filiform papillae are the most abundant projections and are arranged in rows parallel to the sulcus terminalis, except at the apex. They are the only papillae not bearing taste buds (gustatory sense sensors) and are the most keratinized, helping increase friction with food. Larger fungiform papillae occur mostly in the margin and are not keratinized and highly vascular, bearing one or more taste buds. Foliate papillae are found near the sulcus terminalis in four to five lateral folds that are vestiges of larger papillae found in other mammals, none presenting keratinization and having numerous taste buds. Circumvallate papillae are cylindrical 1-2 mm walled structures immediately in front of the sulcus terminalis (8-12 of them) each bearing ~250 taste buds.

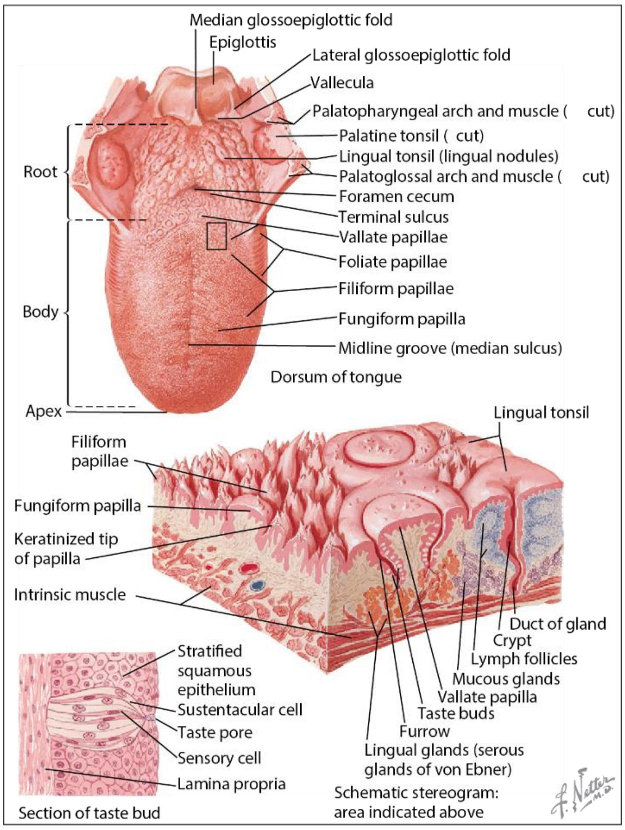


Figure 13 Schematic representation of the distribution and types of lingual papillae on its dorsal surface. From: Ten Cate's Oral Histology Development, Structure and Function (*126*).

There is a total of 32 teeth in the complete adult permanent dentition, eight in each jaw quadrant (right/left maxillary/mandibular). This set replaces the deciduous dentition, infant teeth erupting from six months to three years of age. The permanent set is completed at 18-21 years of age after the eruption of the third molars (which are commonly surgically removed; Figure 14).

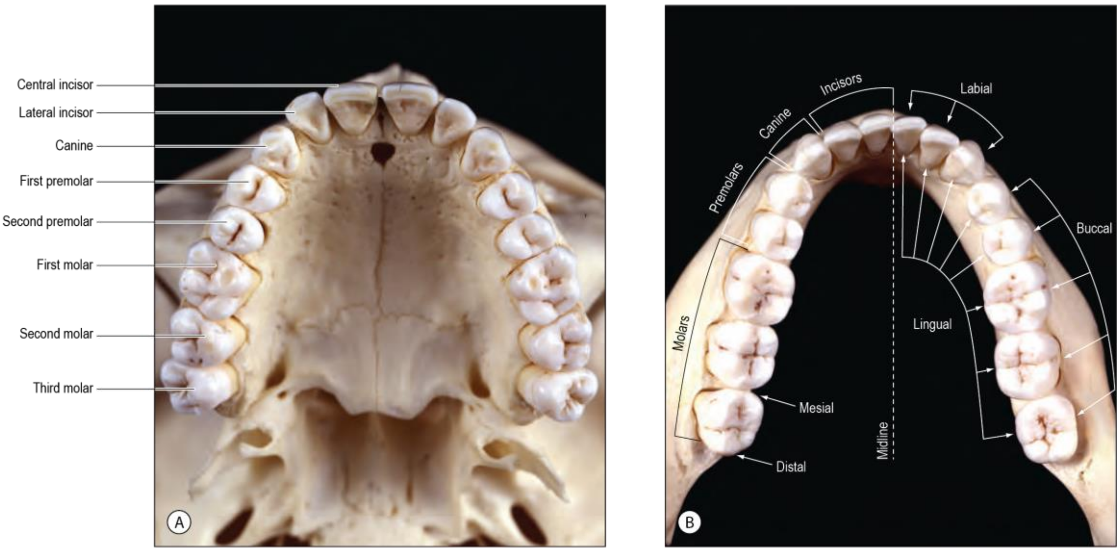


Figure 14. The permanent teeth: occlusal aspect. A, The upper dental arch. B, The lower dental arch. The terminology employed for the identification of teeth according to their location is shown in the lower jaw; the same terminology is used to describe the teeth in the upper jaw. From: Gray's Anatomy 41th Ed (*125*).

Externally, the tooth is composed of a crown covered by hard 2-2.5 mm enamel and a root covered by bone-like cementum, joined by the cervical margin (neck). Internally, the dentine surrounds a central pulp cavity connected to a pulp chamber in its coronal end (Figure 15). Within the root, the same canal narrows and opens at the tip by an apical foramen. This is occupied by the pulp, a soft connective tissue protruding from the periodontal ligament containing vessels and nerves that provides nutritive and immunological support (it has dendritic antigen-presenting cells). The pulp recedes with age, eventually leaving the crown entirely.

The root is surrounded by 0.2mm periodontal ligaments and is attached the alveolar bone with bundles of collagen fibres. The periodontal ligaments support the teeth and provide sensory information and irrigation.

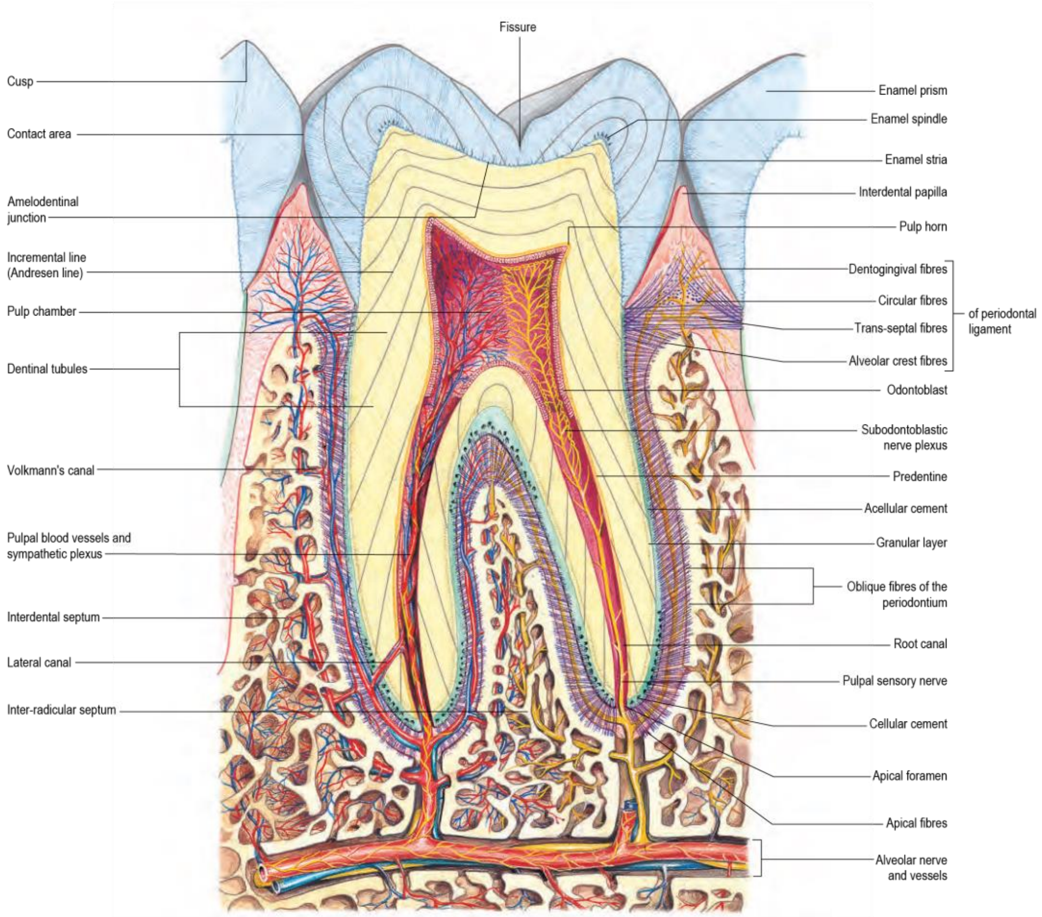


Figure 15 A longitudinal section of a tooth and its surrounding tissues. From: Gray's Anatomy 41th Ed (*125*).

Histology of the oral mucosa

Microbiota colonizes the vast majority of the highly variable surfaces in the oral cavity. Most microbes reside in the mucosa that is in contact with nourishment and has very distinct tissue configurations depending on the location, the mechanical stress it is subjected to and its general function. The following section will contain summarized descriptions from the book *Ten Cate's Oral Histology Development, Structure and Function* (*126*).

The oral mucosa does not present appendages and has salivary and sebaceous glands (yellowish Fordyce's spots). The softer lining mucosa gapes when surgically incised and disperses fluid but the firmer mastical mucosa does not, making it robust but susceptible to painful inflammation.

The two main components of the oral mucosa are the stratified squamous epithelium and the lamina propria, the underlying connecting tissue. The irregular interface is presented with connective tissue papillae (upward projections) and epithelial ridges called rete processes or pegs (downward projections), as well as basal lamina.

In the cheeks, lips and parts of the palate, a loose layer of submucosa lies underneath, having fatty and glandular connective tissue harbouring major blood vessels and nerves. Its separation is not as clearly defined as it is in the gut, where a layer of smooth muscularis mucosae splits the two regions (Figure 16).

Contrastingly, in the gingiva and parts of the hard palate the mucosa presents as mucoperiosteum, as it is attached directly to the periosteum of the bone beneath, providing a firmer grip with no submucosa. Additionally, some lymphoid tissue nodules exist in the oral mucosa, especially in the posterior part of the cavity (tonsils called Waldeyer's ring), as crypts (invaginations of the epithelium into the lamina propria), containing lymphocytes and plasma cells.

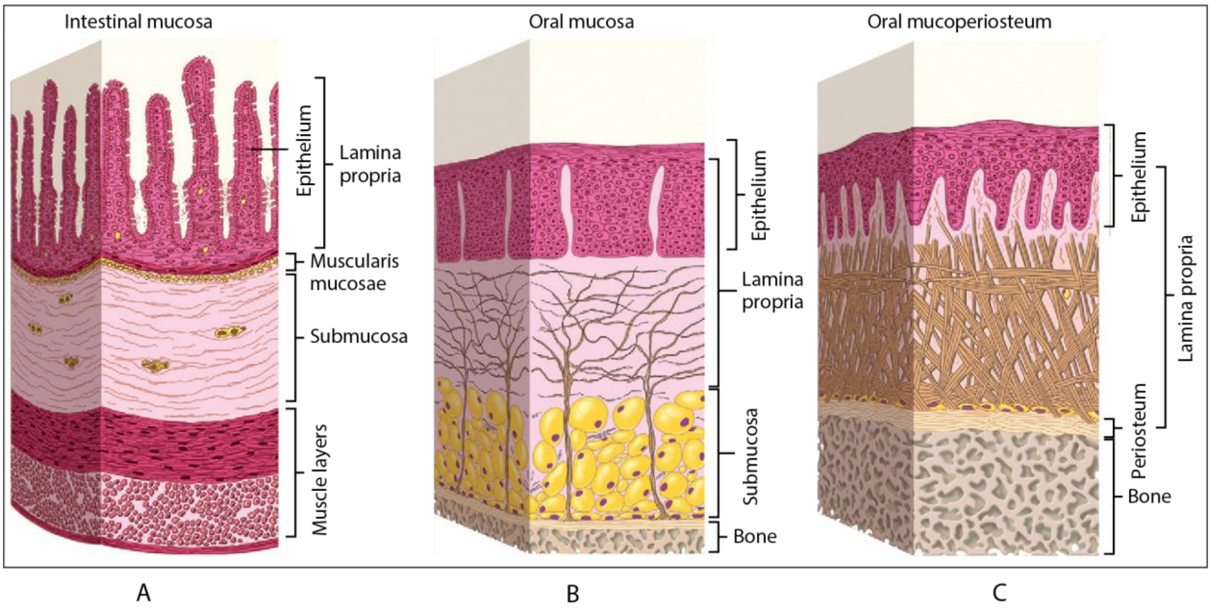


Figure 16. Arrangement of tissue components. A) Intestinal mucosa. B) Oral mucosa. C) Oral mucoperiosteum. From: Ten Cate's Oral Histology Development, Structure and Function (*126*).

Cells in the oral epithelium constitute the primary barrier against external agents, packing in a tight conformation of squamous layers (strata) which are continually renewed by mitotic divisions from the progenitor population in the deepest layers, maturing and differentiating in the outer layers into keratinocytes and later corneocytes. These replace their plasma membrane with a protective layer, the cornified cell envelope, which consists of keratins and other proteins embedded in an insoluble matrix of lipids that, pack together with modified desmosomes (cellular attachment structures), same which are proteolytically degraded to allow eventual desquamation.

In the gingivae, a whole renewal cycle of the epithelium takes 41-57 days while in the cheek it takes 25 days. The main difference arises from the type of maturation, which can be keratinized or non-keratinized. A histological section of mucosa displays four epithelial regions. The first two are always the same: the basal layer (adhered to the basal lamina, where most cells divide) and the prickle cell layer (containing elliptical cells with tonofibril bundles). In mastical mucosa, the third one is the granular layer and is made up of squamous cells (producing kerathohyaline granules which start keratinization). The final, keratinized layer presents a fully dehydrated and extremely flattened cell conformation (Figure 17), which may conserve their nuclei (parakeratinized) or not (orhokeratinized). In lining mucosa (non-keratinized), squamous cells are larger and have only dispersed filaments in its two upper layers, the intermediate and superficial strata. The last one has cells with fewer organelles but nuclei are still present.

Contrary to the skin, parakeratinization of the oral mucosa does not imply disease. Other cells that may be present in the epithelium include melanocytes (pigment producing) and Merkel cells (pressure sensors) in the basal layer and Langerhans cells (antigen trapping and processing) in the suprabasal layers and lymphocytes (inflammatory associated) in various layers.

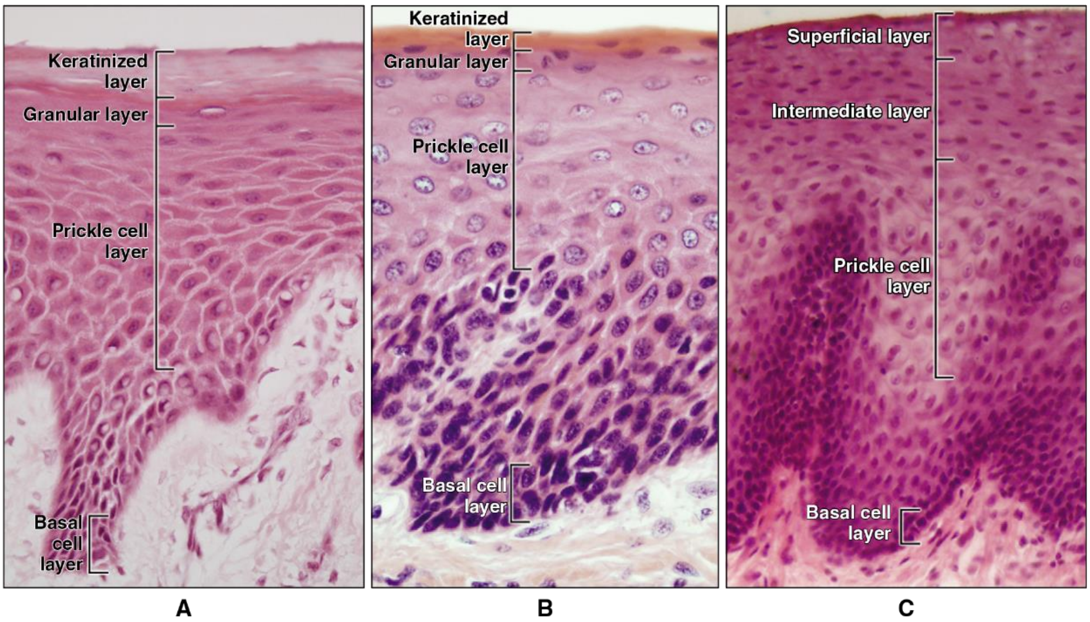


Figure 17. Histological sections of the main types of maturation in human oral epithelium. A) Orthokeratinization in the gingiva. The keratinized surface layer cells have no nuclei. B) Parakeratinization in the gingiva. The keratinized cells in the surface layer retain their nuclei. C) Non-keratinization in buccal epithelium. Cells are thick and present nuclei. From: Ten Cate's Oral Histology Development, Structure and Function (*126*).

There are two layers of the lamina propria (Figure 18), which supports the oral epithelium: the papillary layer presenting papillae associated with the epithelial ridges and the reticular layer, filled with collagen fibres. The lamina is mostly composed of fibroblasts which excrete fibres, only proliferating rapidly as response to wounds. Other cells include macrophages and its precursor histiocytes, mast cells (secreting inflammatory mediators and vasoactive agents), neutrophils, lymphocytes, plasma cells (synthetizing immunoglobulins) and endothelial cells (from vascular channels). In masticatory mucosa, the lamina propria is directly attached to the periosteum (mucoperiosteum) or the muscle (tongue).

The dentin and pulp exist as part of a histological complex within the teeth. The hard dentin forms the bulk of the tooth, consisting of an array of tubules (dentinal tubules) traversing its entirety. It is in constant maintenance by odontoblasts in the outer layer of the pulp via cytoplasmic projections called odontoblast processes. Odontoblasts form dentin prior to the development of the enamel but, contrary to the latter, dentin continues to form through life.

The dental pulp (Figure 19) is formed by four zones: the odontoblastic pulp periphery, the cell-free zone of Weil, the cell rich zone and the core. Its soft connective tissue is mostly composed by odontoblasts, fibroblasts, undifferentiated ectomsenchymal cells, macrophages and other immunocompetent cells.

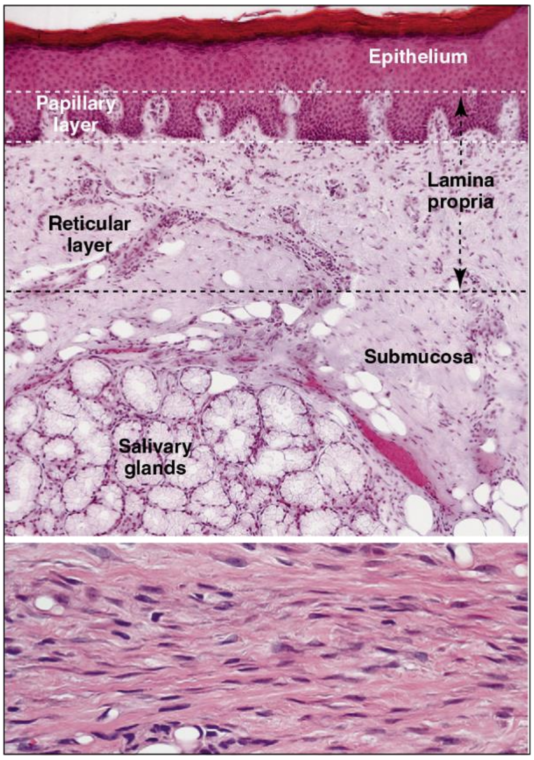


Figure 18. Photomicrograph of palatal mucosa. The approximate boundaries of the papillary and reticular layers are shown. Minor salivary glands are apparent in the submucosa From: Ten Cate's Oral Histology Development, Structure and Function (*126*).

The dentogingival junction (Figure 20), where the oral mucosa meets the tooth, is of particular importance as it harbours bacteria and is subject to inflammation. The healthy gingival sulcus is 0.5-3mm deep while larger sulci called periodontal pockets are considered pathological. This subgingival area migrates gradually with age, normally found in the cervical area of the tooth on adults and reaching the cementum in elders. Polymorphonuclear leukocytes continually get from the epithelium to the sulcus. The oral sulcular epithelium is nonkeratinized, as opposed to that in the free gingiva. Both tissues are continuous and meet at the gingival crest.

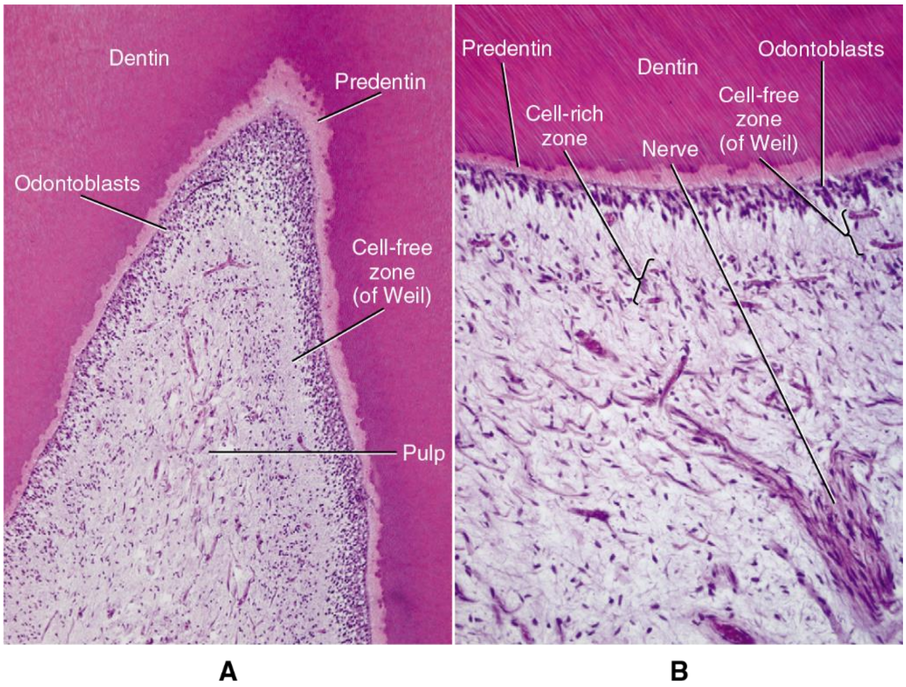


Figure 19 Dental pulp A) Photomicrograph of the dentin-pulp complex. B) At more detail, the different zones of the pulp and the dentin tubules become visible. From: Ten Cate's Oral Histology Development, Structure and Function (*126*).

Deeper in the sulcus there is a thin squamous stratified junctional epithelium. The basal cells deriving the epithelium rest on a basal lamina interfacing with subjacent dermal connective tissue (outer basal lamina). The gingiva is attached to the tooth by a structural complex called epithelial attachment, in which a superficial basal lamina is formed and maintained by squamous cells. The lamina binds to calcified surfaces while cells remain attached to it by hemidesmosomes. The junctional epithelium is incompletely differentiated and its cells have fewer tonofilaments than regular epithelial cells. It has a high rate of cell division forming a 2-3 cell layers on the tooth surface.

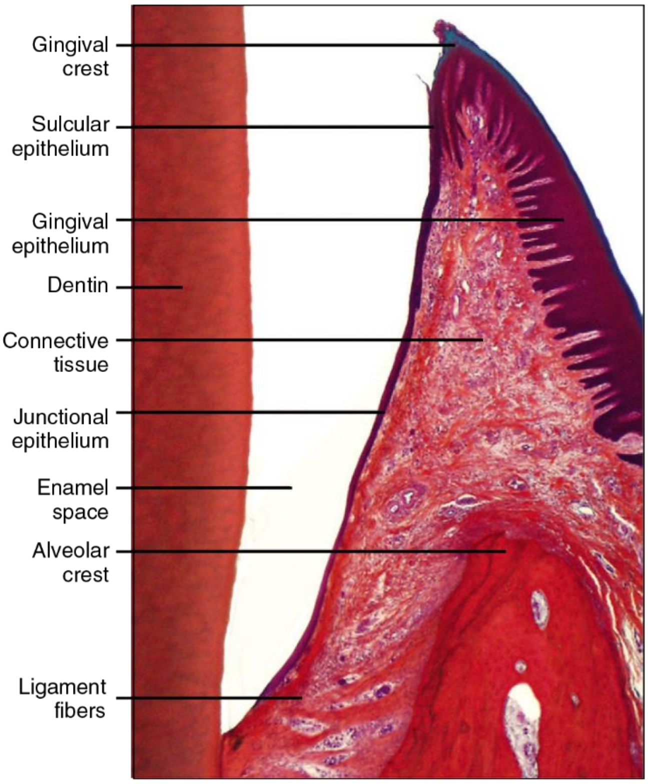


Figure 20.- The dentogingival junction attaches to the tooth surface. In this section, the enamel was removed by decalcification, leaving an enamel space. The sulcular epithelium is short because the tooth is not fully erupted. From: Ten Cate's Oral Histology Development, Structure and Function (*126*).

The oral cavity as an ecological niche

The different niches in the human mouth have very distinct properties for bacteria. The following descriptions were summarized from *Oral microbiology 5ed*. (*127*) unless otherwise indicated.

As a habitat, the oral cavity has a stable temperature (35-36ºC, reaching 39ºC during gingival pockets inflammation) and oxygen concentration (around 20%). Still, the majority of the oral species inhabiting it are either facultatively anaerobic (able grow in presence or absence of oxygen) or obligately anaerobic (requiring an oxygen deprived condition; oxygen can be toxic for them).

Mucosal colonization is restricted by desquamation and some surfaces have specialized host cell types, making it inhospitable for most microorganisms. Keratinization and overall tissue roughness is also a factor affecting its occupation. The tongue, due to its papillated surface and its low redox potential, acts an especially prominent reservoir for anaerobes. Teeth are especially important for bacteria as a niche as they are stable surfaces (not shedding like the mucosa), allowing for the accumulation of large communities in biofilms (dental plaque), especially at protected surfaces such as fissures, interdental papillae and gingival crevices embedded in nutrient-rich gingival crevicular fluid (GCF), providing optimal conditions for the growth of anaerobic bacteria. Dental restorations, crown and bridgework, removable prostheses and implants constitute additional surfaces for bacteria to grow. Also, recession of the gingival tissues resulting from natural aging exposes the cementum to microbial colonization.

Biofilms are a highly organized structural entities composed of aggregated microorganisms (representing 15-20% of their volume) embedded in a intercellular matrix (other 75-80%) normally seen in the teeth as dental plaques appearing over the enamel and gingivae (*128*). Biofilm formation occurs in a progression of aggregation steps and confer the organisms a reported 1,000-fold increase in tolerance to antimicrobial agents when compared to free living cells (Figure 21). During plaque formation, phylogenetically distant bacteria attach to each other using adhesins that recognize protein glycoprotein of polysaccharide receptors on the oral surfaces or bacterial cells. If left undisturbed, a 2-3 weeks plaque grows 50-100 µm thick (*129*). Plaque that mineralizes with calcium and phosphate forms a calculus. Although the process is natural, its increase is related to caries formation and periodontal diseases.

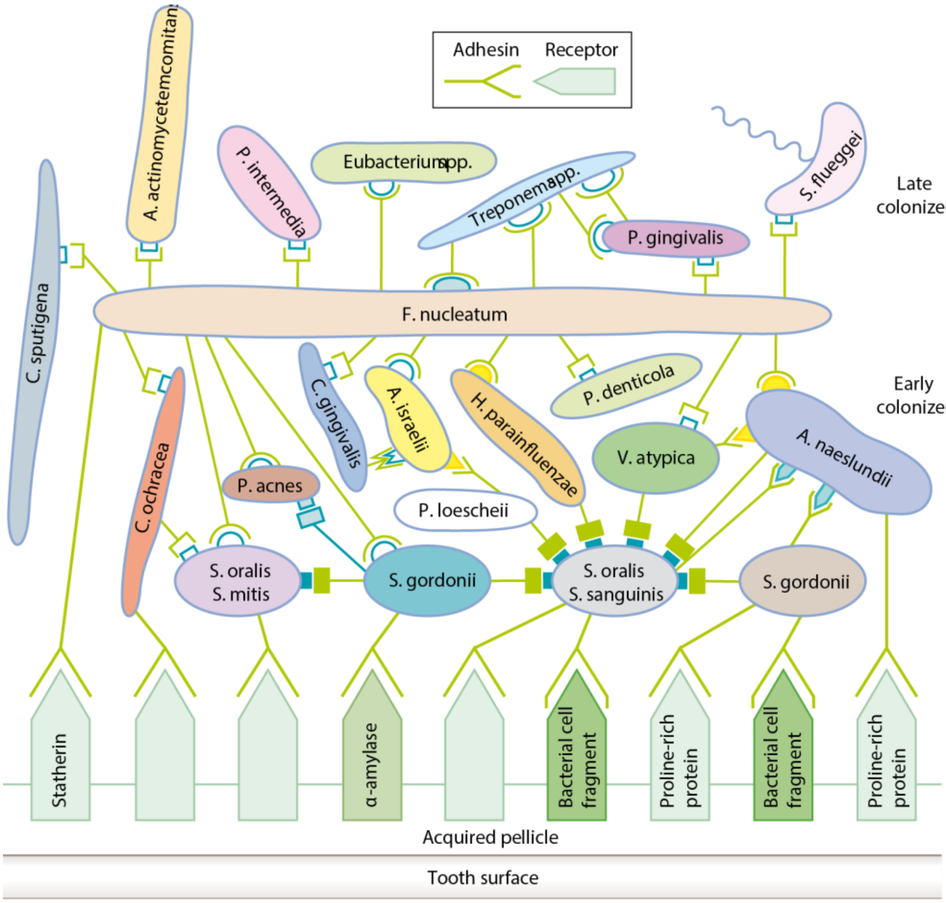


Figure 21. Schematic representation of the patterns of coaggregation in human dental plaque. Early colonizers bind to receptors in the acquired pellicle. Subsequently, early and later colonizers bind to receptors on the surface of these already attached cells using them as scaffolds. Excreted intracellular matrix produced by the bacteria supports the structure. From: Oral Microbiology (*127*).

Microbes also shape its environment by consuming oxygen and releasing carbon dioxide and hydrogen, further contributing to the sustainability of an anaerobic environment. Also, when appear and reach the dentine, the pH decreases due to carbohydrate metabolism, whereas it increases with peptide metabolism. The new conditions dictate shifts in colonizing microbial species at each stage.

Saliva lubricates the mucosa and buffers the pH of potentially damaging acids to stabilize it around 6.75 - 7.25. It also affects biofilm formation and has antimicrobial factors such as lysozyme, lactoferrin and the salivary peroxidase, immunoglobulins and antimicrobial peptides. Furthermore, saliva aggregates exogenous organisms for swallowing and provides nutrients in the form of proteins and peptides. GCF contains serum components that reach the mouth through the junctional epithelium, increasing 30-fold in gingivitis. GCF flushes away non-adherent microbes and provides IgG and neutrophils. It also acts as a source of nutrient for mutualistic bacteria that metabolise proteins. Gingival diseases increment pH (>7.75) in periodontal pockets (abnormally large crevices) promoting the colonization by specialized pathogens.

A broad range of microbes result from differential tolerance to oxygen concentration and pH, particularly in biofilms. Sugar consumption lower the pH in dental plaque below 5.0 mostly by lactic acid production, favouring colonization by acid-tolerant streptococci and Lactobacilli, normally in low abundances and considered cariogenic. On the other hand, high pH during inflammation in periodontal diseases favours the colonization by pathogenic species such as *Porphyromonas gingivalis* and other periodontal pathogens.

Most native microbiota nourishes from endogenous nutrients provided by the saliva and GCF including sugars, glycoproteins, amino acids, mucin, albumin, etc. Exogenous dietary nutrients include carbohydrates that can be converted by bacteria into exopolymers used to consolidate attachment or as extracellular nutrient storage. Protein-rich products can contribute to increase pH due to decarboxylation of amino acids and casein whereas dairy products containing casein can form an adsorbed pellicle reducing adhesion of certain pathogens. Nitrate in green vegetables is converted by bacteria to nitrite, inhibiting other pathogenic bacteria by the conversion to nitric oxide. Fluoride, present in most toothpastes, incorporates into the enamel, preventing its demineralization and inhibits bacterial glycolysis under acidic conditions.

Exploring the culturable oral microbiota

Culture-independent techniques brought a new paradigm to the understanding of the human oral cavity, changing for good a field that is deeply rooted on the development of dentistry, modern medicine and classical microbiology during the 20th century. Dentistry had been traditionally focused on the technical and physiological aspects of the practice, a view that was challenged by the American biochemist William Gies, founder of the first department of biochemistry in a medical school (*130*). As a scientist, Gies studied dental caries and periodontal diseases from a molecular perspective and advocated for an experimentation-oriented university-based education in dentistry on par with medical schools.

Gies' research inspired the next generation of oral research scientists, most notably Theodor Rosebury, to focus on the microbiological aspect of dental caries and periodontal diseases. The oral cavity had become the most studied “microflora” as it was readily accessible. In 1939, Rosebury's research led to identification of carbohydrates in westernized diets as the main source of caries by comparing the diet of Eskimos in Alaska (*131*). Although James Kilian had isolated and characterized *Streptococcus mutans* (now known as one of the important players in caries formation) in 1924 from caries samples (*132*), Rosebury kept on exploring oral microbes, effectively dedicating his career to investigating the relation of nutrition and immunological stress of the microbiota to health and disease.

Rosebury's work in the Columbia University has proven critical in the understanding of the oral cavity as a complex ecosystem. Some of his most important findings include the role of *Lactobacillus* in enamel demineralization and acid production in rat models (*133*) and his collaborations with Genevieve Foley on studies of the relation of fusospirochetal infections and Acute Necrotizing Ulcerative Gingivitis (ANUG, formerly known as Vincent's Infection), a non-contagious affliction presenting bleeding gingivae, severe ulcerations, a fibrinous exudate and foul odour in the interdental papillae (*134*). They proved transmissibility of the infection in guinea pig and rabbit models (Koch’s postulates mandate this for classical demonstration of a microorganism as causative agent). In a key review article, Rosebury noted the inoculation of fusospirochetas in animals produced a simplification of the initial microbiota by the elimination of some of the non-pathogenic microorganisms and proliferation of pathogenic species of spiroform bacilli, vibrios and streptococci (*135*). He speculated this pathology might not transmit in humans because the etiological agents were in fact part of normal microbiota and thus the onset of the disease depended on conditions being optimal (nutritional or immunological deficiency) for them to thrive and trigger it, an idea which was revolutionary at the time. Eventually Rosebury and Foley would sustain four pathogens were the etiological agents of ANUG: a spirochete, a fusiform bacillus, a vibrio, and an anaerobic streptococcus, indicating these were interdependent and together could produce the onset of the disease if overgrown (*136*), later suggesting a whole community of other organisms may procure their anaerobial and nutritional requirements (*137*). His work on spirochetes also led him to develop the first anaerobic chamber (*138*).

Box 4 – Heirs of Rosebury

Rosebury’s influence was profound in the academic community and was understandably considered to be the grandfather of modern oral microbiology (*139*). In the 60s, he co-wrote *A Textbook of Oral Pathology*, an important textbook in the field, was editor of the Journal of Dental research in the 70s, worked on an analysis for the military on microbial warfare that would eventually backfire when made public, forcing him to halt his pioneering work on the first anaerobic chamber for spirochetes (some of these ideas were used by the now standardized Rolf’s Freter chamber). He also wrote two books on the topic of human microbiota: the textbook Microorganisms Indigenous to Man (1962) and popular Life on Man (1969).

But more importantly, Solon Ellison and Jack MacDonald, two of Rosebury's students passed on the knowledge to some of the most influential oral scientist of the next generation (*139*). Ellison left to SUNY Buffalo, New York, to fund an Oral Biology department where he created the first PhD program in Oral Biology forming Drs. Baker, Chen, Crone, Emmings, Haraszthy, Herzberg, Levine, Macvittie, Miyasaki, Patters, Ramasubbu, Reed, Reddy, Scannapieco, Schenkein, Schifferle, Tabak, Taubman, Van Dyke, and Zambon. MacDonald would become Director of the Forsyth Dental Infirmary for children (now Forsyth Institute, one of the leading centres for dental research worldwide) recruiting Sigmund Socransky and Ronald Gibbons, forming Drs. Bratthall, Caufield, Clark, Dewhirst, Ellen, Haffajee, Hillman, Kelstrup, Liljemark, Listgarten, Loesche, Offenbacher, Paster, Paunio, Taichman, Tanner, Stashenko, Slots, Walker, and Williams.

Institutions such as the Harvard Medical School of Dental Medicine and its affiliated Forsyth Institute, the National Institute of Dental Research, as well as the Faculty of Odontology in the University of Gothenburg and the Royal Dental College in Denmark (as part of the Scandinavian scene), generated some of the most important developments in oral research over the second half of the 20th century. In 1960, Paul Keyes proved infection and transmissibility in hamsters of a streptococcus proving its causality in cariogenesis (*140*). In 1965, Jepsen and Winther, suggested a possible role of the yeast-like fungus *Candida albicans* in infections of erythroleukoplakia (speckled leukoplakia), a non-homogeneous lesion presenting white keratotic patches in red atrophic mucosa, most commonly detected in the commissural area (*141*). They also demonstrated antifungal treatment leads to a change into homogeneous leukoplakia (well defined white patches). In a paper by Winner in 1969, he studied Candida's switch from commensalism to parasitism, noting the absence of other flora as a possible contributing factor (*142*). In 1973, and working on rat model, Russel and Jones confirmed it as causative agent of infections causing the appearance of parakeratotic patches in the tongue, making it lose its papillary structure, a pathology now known as oral candidiasis, after achieving infection by eliminating bacterial competitors with a prolonged antibiotic treatment (*143*).

Sigmund Socransky, recruited to the Forsyth Institute by Jack MacDonald (once mentored by Rosebury; see Box 4) became one the most influential figures of that period, focusing on localized aggressive, chronic and refractory periodontitis (diagnosed when patients fail to respond to at least three forms of therapy) (*144*). In 1964, Socransky and collaborators. obtained the first pure culture of spirochetes, the conclusion of a long series of experiments that first identified diphtheroids and fusobacteria as necessary accessory organisms to grow spirochetes *in vitro* and then extracted the necessary growth factors they produced, putrescine and isobutyrate (*145*) further describing and characterizing *Treponema denticola*, a strict anaerobe spirochete with important proteolytic activity as one of the etiological agents of periodontitis (*146*). Socransky was also interested in the study of ANUG but the incidence of the disease suddenly fell close to zero during the mid-60s (*144*).

As many of his colleagues, Socransky worked on the microbial complexes of the subgingival plaque, identifying pathogens and host-compatible species, including descriptions of different vibrio (*147*), and bacteroides (*148*), and is responsible, over his collaborations with Gibbons, MacDonald, Listgarten and Tanner, for naming several human oral commensals and pathogens such as some *Treponema*, *Bacteroides*, *Capnocytophaga*, etc. (*148–150*), etc. The oral research community were early adopters of DNA-based methods such as the DNA-DNA hybridization after its publication in 1970, a method that measures relatedness amongst bacteria, based on renaturation rates of DNA given by their G+C contents (*151*). This was quickly reflected in the reclassification of several oral microbes during the next decades in efforts such as a taxonomical revision by Kilian in 1975, of genus *Haemophilus* (*152*), reporting *Agregibacter segnis* (previously *Haemophilus segnis*), an oral pathogen that has been identified as an important pathogen causing bacteremia (*153*). Other examples include the creation of genus *Wolinella* to allocate some formerly vibrio bacteria (*154*) and collapsing of some more *Haemophilus* species (now *Agregibacter*) as they were not distant (*155*).

In 1976, Listgarten used electronic microscopy to analyse tooth surfaces in the presence of periodontal health and varying degrees of periodontal disease and determined microbiota was significantly altered in disease (*156*), noting the importance of working with the microbiota as a community that can fluctuate in response to an stimulus. As DNA sequencing standardized, the oral research community adverted the importance of studying the microbiota as a complex habitat. This was made patent by the growing interest in biofilm research (*157, 158*).

Socransky was convinced that only a limited fraction of microbes were pathogenic and some may even be beneficial for its host (*144*). To cope with this, during the early 90s, he and his colleagues improved DNA hybridization approach in a semi-automated approach called checkerboard DNA-DNA hybridization that allowed surveying large numbers of DNA samples with multiple probes of 16S rDNA or whole genomic sequences, much like a microarray with hundreds of fluorescent probes (*159*). The technique popularized in the oral research community as it would enable the study of a whole community of bacteria at a time, allowing quantification and the application of ecology-oriented statistical analysis as was demonstrated in a classical analysis by Socransky and collaborators in 1998 in which they described bacterial clusters in subgingival plaque (*160*).

In 2000, still before the advent of high-throughput sequencing and metagenomics, Tanner and collaborators published an important review article containing the phylogeny of the oral microbiota that was known at the time, using the 16S rRNA gene as phylogenetic marker (*161*). Although this has been much improved over the years and taxonomy has been thoroughly revised, this effort set a framework for the analysis of the bacterial diversity of the oral cavity. Furthermore, they urged the scientific community to explore the unculturable fraction of the oral microbiota.

A long road was travelled by the scientific community doing oral research in the 20th century and this summary by no means intends to be exhaustive but to pay respects to the enormous clinical effort done before and note key moments that summarize the advance of the field towards culture-free studies. However, prior to reviewing these, it is important to name culturable bacteria. To date, more than 460 species from oral niches have been cultivated (*162*) and the most prevalent are clearly identified (Table 1) so far in the oral cavity and point out the most prevalent culturable ones in different niches (Table 2). As of 2017, approximately 50% of the oral species have been cultivated and characterized with a valid, ~10% remained unnamed after being cultivated and the rest are just known thanks to culture-free molecular approaches (*163*).

Table 1. The principal culturable bacterial genera found in the oral cavity (*127*).

Gram-positive bacteria		Gram-negative bacteria	
Cocci/others	Rods	Cocci/others	Rods
Abiotrophia	*Actinobaculum*	*Anaeroglobus*	*Aggregatibacter*
Enterococcus	*Actinomyces*	*Megasphaera*	*Campylobacter*
Finegoldia	*Alloscardovia*	*Moraxella*	*Cantonella*
Gemella	*Arcanobacterium*	*Neisseria*	*Capnocytophaga*
Granulicatella	*Atopobium*	*Veillonella*	
Peptostreptococcus	*Bifidobacterium*	*Centipeda*	
Streptococcus	*Corynebacterium*	*Desulfomicrobium*	
Cryptobacterium	*Desulfovibrio*		
Eubacterium	*Dialister*		
Filifactor	*Eikenella*		
Lactobacillus	*Flavobacterium*		
Mogibacterium	*Fusobacterium*		
Olsenella	*Haemophilus*		
Parascardovia	*Johnsonii*		
Propionibacterium	*Kingella*		
Pseudoramibacter	*Leptotrichia*		
Rothia	*Methanobrevibacter*		
Scardovia	*Porphyromonas*		
Shuttleworthia	*Prevotella*		
Slackia	*Selenomonas*		
Solobacterium	*Simonsiella*		
Tannerella			
Treponema			
Wolinella			

Table 2. Proportions of some cultivable bacterial populations at different sites in the normal oral cavity (*127*)**.**

Bacterium	Saliva	Buccal mucosa	Tongue dorsum	Supragingival plaque
Streptococcus sanguinis	1	6	1	7
S. salivarius	3	3	6	2
S. oralis/S. mitis	21	29	33	23
mutans streptococci	4	3	3	5
Actinomyces naeslundii	2	1	5	5
A. odontolyticus	2	1	7	13
Haemophilus spp.	4	7	15	7
Capnocytophaga spp.	<1	<1	1	<1
Fusobacterium spp.	1	<1	<1	<1
Black-pigmented anaerobes	<1	<1	1	+*
* Sometimes present				

Beyond the culturable

By the end of the 20th century, the oral research community had established reliable methods to grow oral bacteria in pure culture, naming and characterizing many of the numerous pathogens causing major oral diseases in the population. Still, a major fraction of the microbial population eluded cultivation, a fraction that had been suggested as critical to the overall ecosystem understanding. A step forward was the adoption of PCR amplification, to detect and quantify non-culturable species although high-throughput sequencing would arrive relatively late, in favour of cloning full 16S rDNA sequences and checkerboard analyses.

The first of such studies in oral microbiota was published in 1996 by the group of William Wade by sequencing the whole 16S rRNA gene from DNA obtained from three dentoalveolar abscesses (pus) where they found *P. gingivalis* and *Prevotella oris* as well as bacteria from the *Peptostreptococcus micros*, and some uncultured *Prevotella* and *Zoogloea*. Although this was a fairly small experiment, it demonstrated the method’s effectiveness and prompted other groups to join. However, it was not until the 2000s that things really started for unculturable analyses.

One of these was the first non-culturable survey of the composition of the microbiota in the subgingival plaque in which five healthy subjects and 26 subjects with periodontal diseases were analysed (nine with periodontitis, a set of afflictions of the tissues surrounding teeth that lead to inflammation, loss of the alveolar bone and eventual teeth loss, eleven with refractory periodontitis, two with HIV periodontitis, diagnosed HIV+, and 4 with ANUG) (*164*). They found nine bacterial phyla including periodontal pathogens such as *P. gingivalis*, *Bacteroides forsythus*, and *Treponema denticola* in low proportions in all samples, including healthy ones, as well as several species that had not been reported in the oral cavity as well as several new phylotypes, suggesting other bacteria may have a role in the pathologies. The group would revisit subgingival plaque some years later to analyse generalized aggressive periodontitis in ten subjects with the same methods finding a marked prevalence of *Selemonas* and *Streptococcus*, determining the former may contribute to the affliction (*165*). Still, they noted some bias may have prevented *Aggregatibacter actinomycetemcomitans* from being detected.

In other study, Paster and collaborators explored the microbiota associated to advanced noma, a severe gangrenous disease necrotizing large areas of soft and hard tissue of the mouth, cheeks and nose of four subjects in Nigeria (*166*). One third of the 16S rDNA sequences were from uncultured organisms, identifying seven phylotypes unique to the disease samples, some of which were presumably transferred from soil, and *Fusobacterium* spp., previously proposed as pathogen of noma. The same year, Munson and collaborators explored a less described niche, endodontic infections from aspirate samples from root canals of five teeth with 16S rDNA clones (*167*). Firmicutes dominated the niche with *Dialister* as the only bacterial genus found in all samples. The relatively low diversity found was explained by the authors due to few bacteria being able to adapt to conditions in a necrotic pulp, probably due to metabolic constraints. Also, 18 new Firmicutes were found, as well as eight Bacteroidetes.

A different study analysed the microbiota associated with necrotizing ulcerative periodontitis (eight subjects), a painfully affliction affecting 2-6% of HIV-positive subjects (*168*). A total of 108 phylotypes were identified (60% culturable) of which 26 were novel to the pathology whereas common periodontitis pathogens were only found in low quantities. In 2002, Becker and collaborators published the results of the first unculturable assay to study caries in children (30 subjects), to identify other bacteria, besides *S. mutans* that may be associated to caries formation (*169*). Again, several species not previously reported for this niche were detected, some of which were non-culturable. *Streptococcus sanguinis* was associated with health and several other species were associated with caries, including *S. mutans*. A study of halitosis (malodour produced by high levels of volatile sulphur compounds in six of them) in the tongue dorsum of eleven subjects, patients presented higher levels of *Atopobium pavulum*, *Eubacteruim sulci*, uncultured TM7 (now *Saccharibacteria*), among several non-culturable species (*170*). *Streptococcus salivarius* and *Rothia mucilaginosa* were most prevalent in non-affected patients.

Munson and collaborators analysed the unculturable fraction of the caries-associated microbiota by cloning the 16S rRNA gene from five patients with advanced caries via cloning (*171*). This study found the predominant taxa to be *S. mutans*, *Lactinobacillus gasseri*/*johnsonii* and *Lactobacillus rhamnosus*. Proportions of species found and their total numbers differed but the composition at the genus level was similar between them, suggesting that different species could provide similar functionality in the biofilm. They hypothesised the structure of biofilms, nurturing and signalling requirements were the limitations that refrained those species to be cultured. Archaea became part of the non-cuturable picture as one third of a cohort of 50 subjects with periodontitis in a study by Lepp and collaborators, all from a phylotype close to *Methanobrevibacter oralis* first isolated, cultured and characterized in the 90s (*172*) and restricted to the subgingival area, possibly forming a syntrophic relationship with increased sulphur producing *Treponema* populations (*173*). More sensitive studies later found they were in all healthy subjects and found the others to be *Methanobacterium curvum*/*congolens* and *Methanosarcina mazeii* (*174*), hypothesizing a role of archaea in the consumption of hydrogen and carbon dioxide in biofilms, favouring the proliferation of fermentative organisms.

In 2005, Aas and collaborators set to explore different niches to define the healthy human oral cavity biome using culture-independent methods (*175*) using a small cohort of five subjects sampled in the dorsum and lateral sides of the tongue, buccal fold, hard palate, soft palate, labial gingiva and tonsils of soft tissue surfaces, and supra gingival and subgingival plaque from tooth surfaces. They managed to detect 141 predominant species, 60% of which were uncultured. Species from genera *Gemella*, *Granulicatella*, *Streptococcus* and *Veillonella* were reported to exist throughout the oral cavity with *S. mitis* being the most extended and all sites having 20-30 predominant species. From this and several other similar cloning and sequencing studies, Dewhirst and Paster collected oral bacteria rDNA sequences, preparing multiple probes for an upgraded version of the checkerboard DNA hybridization, the microarray-like approach they presented in 1994 along with Socransky (*159*). A membrane of probes made from hypervariable regions of multiple 16S sequences was thus hybridized overnight with pools of amplified rRNA genes from samples, avoiding the need of cloning and sequencing. Although detection would be limited to probes in a given membrane, it would allow the detection of previously identified unculturable species and the cost and effectiveness would prove appealing for clinical researchers, especially for large sets of samples.

One of the earlier takes on this checkerboard analyses was a large-scale survey by Mager and collaborators in 2003. They analysed eight different soft mucosa samples from 225 systematically healthy individuals by DNA-DNA hybridization using 40 bacterial probes of different species (*176*). They noted that species profiles were significantly different depending on the surveyed niche, with the most similar being the dorsal and lateral surfaces of the tongue, the ones in soft tissues and the teeth colonizers. Although restricted to a limited number of species, it was clear which niches had a particular microbial fingerprint. Another large project using the approach was a twin study (n=204, 80 monozygotic and 124 dizygotic) launched in order to investigate the heritability of oral-associated microbiota using 82 probes with checkerboard hybridization. *S. mutans* and several lactobacilli were more abundant in caries-active subjects whereas ten others (especially *S. parasanguinis, A. defectiva* and *G. haemolysans*) were associated to healthy teeth. Heritability estimates were moderate to high for oral species, suggesting there could be a some influenced of the host genome, especially for some hypothesized *S. sanguinis*, *S. salivarius*, and *S. mitis* but were they not able to discard the effect of environmental contributors to the relative amounts (*177*).

In an effort to gather and organize the knowledge about the oral microbiota in the brink of the growing number of 16S sequences available, Dewhirst and collaborators (*124*) published in 2010 the Human Oral Microbiome Database (HOMD), providing curated 16S rDNA gene sequences and corresponding analytic tools to serve as reference for the study of the oral microbiota. This database was constructed from a vast archive of healthy and disease isolates (over 36,000 clones from >1,000 isolates including periodontitis, caries, endodontic infections, and noma) and existing sequences (acute necrotizing ulcerative gingivitis, HIV-associated periodontitis, and ventilator-associated pneumonia). The HMP was a major contributor to this effort as they had a projection of analysing 300 samples from oral niches and the Dewhist's group contributed with sequences and strains (*178*). Their phylogenetic analysis resulted in the observation of 13 different phyla, a number that has increased in 2017 with current figures including 15 phyla (Table 3), 33 classes, 52 orders, 93 families, 199 genera, and 733 species (*163*).

There have been some efforts to culture and study the less prevalent or rare groups in microbiota. One such study was a 16S cloning and fluorescent in situ hybridization (FISH) approach by Vartoukian and collaborators for the characterization of a candidate phylum (now accepted) Synergistetes (*179*). *Jonquetella anthropi* and *Pyramidobacter piscolens* had been cultured before from oral samples but several others remained unnamed and had even been misclassified in previous studies as the authors note. They found 12 of them were more prevalent in periodontitis samples and revealed them to be large curved bacilli by FISH.

Table 3. Current confirmed phyla 16S rRNA sequences. Candidate phyla names are in place (*163*)

Phylum List	Taxa
Firmicutes	258
Bacteroidetes	127
Proteobacteria	119
Actinobacteria	92
Spirochaetes	51
Fusobacteria	39
Saccharibacteria (TM7)	19
Synergistetes	10
Gracilibacteria (GN02)	5
Absconditabacteria (SR1)	5
Chlorobi	3
Chloroflexi	3
Chlamydiae	1
Euryarchaeota	1
WPS-2	1

Although the second generation sequencing technologies had been around for some years by that time, part of the oral research community, mainly that in Harvard/Forsyth, worked with checkerboard assays and cloning instead (*180*, *181*). With a growing collection of 16S probes available (598 in 2016), Paster and Dewhirst launched the Human Oral Microbe Identification Microarray (HOMIM), a service for analysing samples using their checkerboard approach that produced around 53 publications until 2016 when it shifted to current generation sequencing (*182*). In one such study, Colombo and collaborators detected increased species diversity in subjects with refractory periodontitis or treatable periodontitis when compared with clinically-healthy ones. They scanned over 300 species in 67 patients (20 healthy, 30 with treatable periodontitis and 17 with refractory periodontitis) (*180*). The aggressive form of periodontitis presented higher presence of putative periodontal pathogens like *Parvimonas micra*, *Campylobacter gracilis*, and *Eubacterium nodatum*, but also of unusual species, such as *Desulfobulbus* sp. OT 041 (oral taxon number from HOMD) and unculturable TM7 (now *Saccharibacteria*) 346/356. They speculated patients receiving scaling and root planning (mechanical treatment) and antibiotic treatments may succeed in reducing the populations of known pathogens, however favouring resisting unusual species to thrive and carry on with disease due to a change in the ecological conditions. This was further explored in a follow-up study that detected various species that were reduced in good responders to the treatment but not in more aggressive periodontitis, including *Bacteroidetes* sp., *Porphyromonas endodontalis*, *P. gingivalis*, *Prevotella* spp., and *Tannerella forsythia*, among several others (*181*).

The last years of the 2000s would see the first modest but important incursions of metagenome publications about the oral microbiota using high-throughput sequencing. In 2008, the group of Keijser from the Academic Centre for Dentistry Amsterdam (ACTA) published the results of the explorations of the microbiota in healthy adults using pyrosequencing-based 16S profiling (*183*). They collected 71 saliva and plaque samples and analysed the amplicons in a GS-20 sequencer. The outcome was a significant increase in the total number of species in the oral cavity when compared to Paster's highest estimates of ~700 common oral species (*124*) as Keijser detected ~1,000 core species (94% identity clusters because only V6 16S region was used) in that comprised 95% of all clusters with over 6,800 total reported and estimated number of over 19,000. Even so, recruitment graphs suggested sampling was still incomplete. They also reported 318 genera with *Prevotella, Streptococcus*, and *Veillonella* representing 50% of all sequences in saliva and *Streptococcus, Veillonella, Corynebacterium, Actinomyces, Fusobacterium*, and *Rothia* summing up 50% in plaque. Some scepticism is prudent with such large figures as methodological caveats may have influenced the totals (e.g. low abundance reads were not removed and a small 16S region and odd cut-off value for species identity percentage were employed), though it is probable previous estimates had been underestimated. An improved study from scientists at the ACTA, this time with a GS-FLX pyrosequencing platform and samples from dental surfaces, cheek, hard palate, tongue and saliva in three individuals aimed at defining the healthy core (*184*). They found ~500 species (this time they used only clusters with at least 5 sequences) much of which were evenly distributed among the different sites with more marked differences between mucosa, saliva and teeth. A different study, using Illumina technology (GAII), was carried out by the Lazarevic and collaborators and published in 2009 (*185*). This was much smaller in scale, with only three healthy individuals providing saliva and oropharyngeal samples to the study but was restricted to classifiable sequences with the RDP database. For the most part, results were in accordance with Keijser's, showing that Firmicutes, Proteobacteria, Actinobacteria, and Fusobacteria were the main phyla but Bacteroidetes was virtually missing in comparison, something that was explained by possible natural bias in sampling or from the extraction procedure (for the region in the 16S was expected to recover most of them). A total of 135 genera were reported (*Neisseria* and *Streptococcus* constituted 70% of all sequences) and over 8,000 species but it is difficult to compare them to Keijser's results as Lazarevic implemented more stringent filters (e.g. in sequences appearing <=3 times or with >30% more divergent than the closest reference were removed). Still, differences may be explained by the varying sampling areas and the 16S region employed.

The publication of the HMP results in 2012 marked a major landmark in the understanding of the oral microbiota as a community, shedding some light into less explored areas of the unculturable bacteria both taxonomically and functionally (*84*). Apart from being the largest non-culturable oral study up to that point (>310 samples per location) focusing on multiple well-defined niches (cheek, hard palate, keratinized gingiva, palatine tonsils, saliva, subgingival and supragingival plaque and oropharynx) it also included a large set of truly metagenomic WGS sequences (1-7 samples from most sites and >100 from buccal mucosa and supragingival plaque) for functional analyses (*82*).

The subgingival plaque presented the most different species in the oral niches (Figure 22) with over 1,500 species (200 samples recruited), followed by supragingival plaque and palatine tonsils, then the tongue dorsum, saliva, and much below, the buccal mucosa and hard palate, then the keratinized gingiva (recruiting approximately 500 species at 200 samples) (*82*). All were below the total gut in total numbers. Also, the total reported number of genes detected from microbiota inhabiting the buccal mucosa was more than 10-fold that in the supragingival plaque, suggesting that although diversity is larger in the plaque, these bacteria have mostly the same functions whereas there may be more survival mechanisms at work in the cheek.

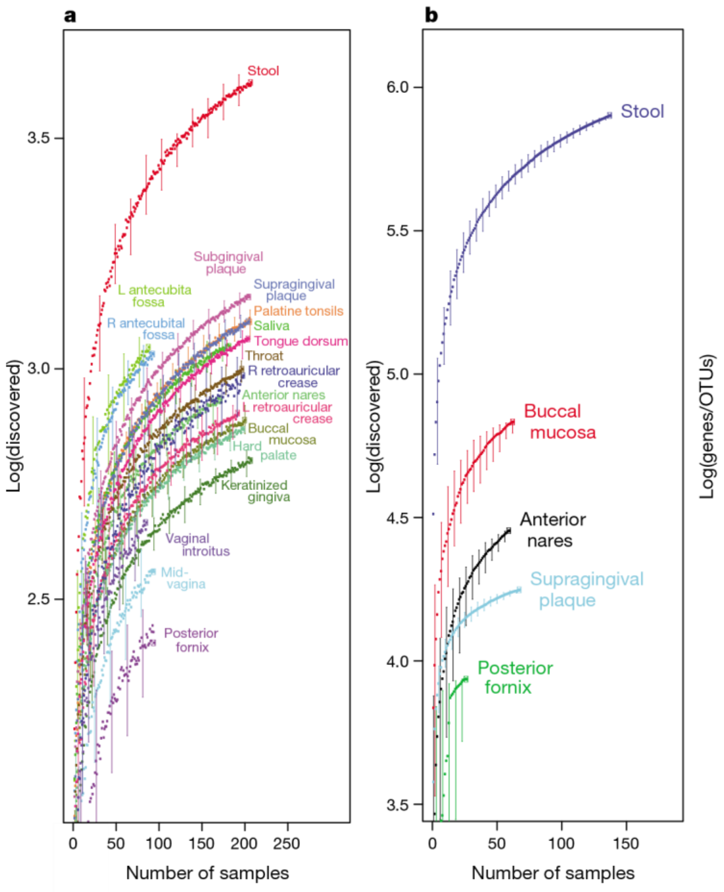
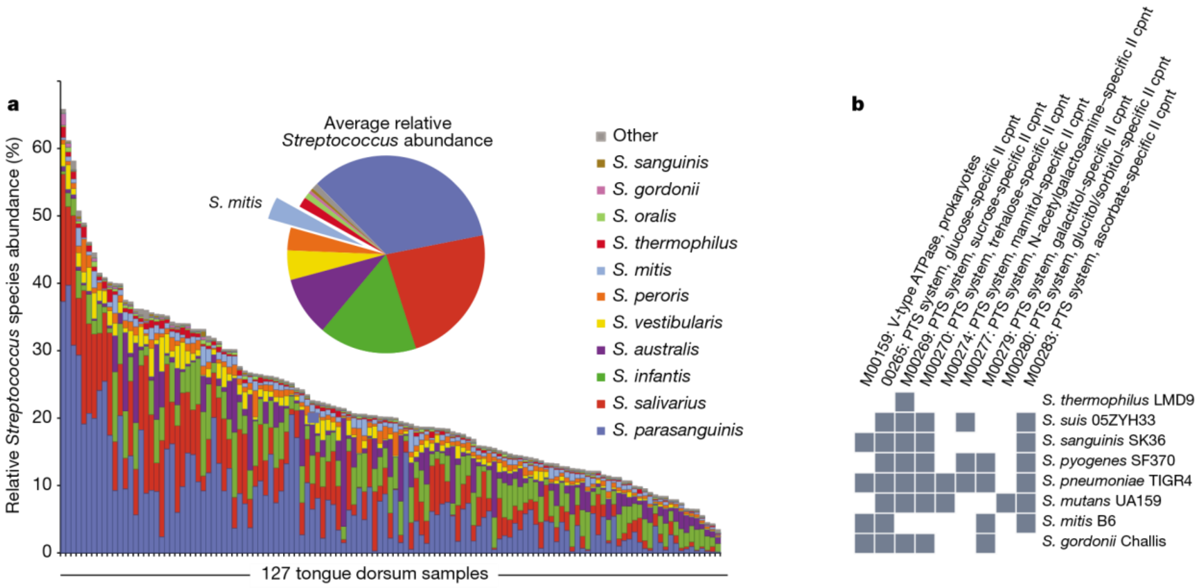


Figure 22. Recruitment plots with no replacement of HMP data. A) Accumulation curves 97% identity clusters (predicted species) of different niches from the HMP data. B) Clustered gene counts from sites with WGS data. From: Methé *et al.*, 2012 (*82*).

The HMP also reported that when all data were considered, oral microbiota was the most stable between subjects and that oral clusters were clearly differentiated from the rest in terms of variability between subject (*97*). From this study, genus *Streptococcus* was dominant in most subjects but it was followed in abundance by *Haemophilus* in the buccal mucosa, by *Actinomyces* in the supragingival plaque and by *Prevotella* in the subgingival plaque. To analyse the extent of the impact of these variations, the consortium further analysed the *Streptococcus* genus in the tongue, demonstrating a large inter-individual variation of the species (Figure 23). This was shown to be functionally relevant as pathways were differentially present in different reference genome strains, thus indicating variation at the interpersonal level appears to be widespread and functionally relevant.

In a deeper taxonomical report by the HMP consortium (*186*), the shifts at the phylum level in the subgingival plaque were reported to be driven by oxygen availability as there was an increase in anaerobic genera *Fusobacterium*, *Prevotella*, and *Treponema*, as well as decreased *Dialister*, *Eubacteruim*, *Selenomonas* and *Parvimonas*. Contrastingly, groups increased in the supragingival plaque including facultative anaerobic genera *Streptococcus*, *Capnocytophaga*, *Neisseria*, *Haemophilus*, *Leptotichia*, *Actinomyces*, *Rothia*, Corynebacterium and *Kingella*. Common pathogens like species of *Porphyromonas*, *Treponema* and *Tannerella* were present in most oral sites but lower in keratinized mucosa, suggesting they are normally part of the commensal microbiota. Also, the less studied *Synergistetes*, TM7 (*Saccharibacteria*) and SR1 (*Absconditabacteria*) phyla were confirmed to occur in the samples.



M00159: V-type ATPase, prokaryotes	00265: PTS system, glucose-specific II cpnt	M00269: PTS system, sucrose-specific II cpnt	M00270: PTS system, trehalose-specific II cpnt	M00274: PTS system, mannitol-specific II cpnt	M00277: PTS system, N-acetylgalactosamine–specific II cpnt	M00279: PTS system, galactitol-specific II cpnt	M00280: PTS system, glucitol/sorbitol-specific II cpnt	M00283: PTS system, ascorbate-specific II cpnt	
		■							*S. thermophilus* LMD9
	■	■	■		■			■	*S. suis* 05ZYH33
■	■	■	■					■	*S. sanguinis* SK36
	■	■	■		■	■		■	*S. pyogenes* SF370
■	■	■	■	■	■	■		■	*S. pneumoniae* TIGR4
	■	■	■	■			■	■	*S. mutans* UA159
■	■					■		■	*S. mitis* B6
■	■	■	■			■			*S. gordonii* Challis

Figure 23. Microbial carriage varies between subjects down to the species and strain level. a) Relative abundance of 11 *Streptococcus* spp. from 127 tongue samples and average composition. b) Present (grey) metabolic pathways in reference genomes of streptococci strains. From: Huttenhower et al., 2012 (*97*).

The oral samples showed the oral cavity had the most diverse protein families in all the HMP dataset totalling ~400,000 protein families, 58% of which were uncharacterized (still, the lowest percentage in al niches) (*97*). Regarding pathways, in general, a large variation was detected in the Sec and Tat secretion systems, suggesting a high degree of host-microbe and microbe-microbe interactions. Specifically, there was a high variability in phosphate mono- and di-saccharide and amino acid transport in the mucosal microbiota and lipopolysaccharide biosynthesis and spermidine/putrescine synthesis and transport on the plaque and tongue. Some more specific functional analysis from the HMP consortium (*186*) indicated that bacteria from the oral niches were detected to be enriched in transporters for small sugars mannose, fructose and galactosamine. The Supragingival plaque was enriched for trehalose, alpha-glucosides and cellobiose transport. Also, putrescine transporters were prevalent in the oral sets. The tongue and supragingival plaque were enriched in iron transporters, linked to an increase in *Porphyromonas* and *Prevotella*. Also, utilization and hydrogen production in the oral cavity were linked to genera *Veillonella* and *Selenomonas* and to an unclassified Pasteurellaceae clade in the supragingival plaque and tongue.

Different longitudinal studies show that the oral microbiota appears to be stable over time, as demonstrated in a pyrosequencing 16S profiling study by Lazarevic and collaborators (*187*). They found for five adults that it was fairly stable at the genus level over three time points over 5 days despite there being intra-individual variation. Similarly, a study of 10 healthy subjects and 26 sampled sites showed inter-individual variation was found to be relatively stable at the genus level in mostly all niches. Stability was also reported between individuals over different geographical regions, as showed in a study by Nasidze and collaborators, who analysed 120 individuals (saliva) from a total of 12 locations worldwide with a 16S cloning approach. The low overall variance in genera composition supported food was not a major driver affecting diversity as most such nutrients and organisms are transitional and main nourishment in for oral bacteria come from saliva and GCF. Variation at the species and strain level was hypothesised to be ecologically redundant (*188*), supporting the findings by Munson and collaborators in 2004 (*171*).

Contemporary to the publication of the HMP results, other groups focused on studying oral diseases with high-throughput techniques. In a taxonomic comparative study using 16S profiling (pyrosequencing), of the subgingival communities of 22 subjects with chronic periodontitis (*189*), defined the core taxonomic groups for disease bearing samples versus healthy samples. Periodontitis samples had a higher prevalence of some spirochetes, *Synergistetes, Firmicutes* and *Chloroflexi*, and lower of actinobacteria, particularly *Actinomyces* when compared to healthy controls. In a complex disease as periodontitis, an important observation is that most disease-associated taxa were also present in controls, meaning that changes in structure may be due to large ecological changes. These results were congruent with a contemporary study by Griffen and collaborators on 29 chronic periodontitis subjects and 29 controls (*190*), supporting *P. gingivalis*, *T. denticola* and *T. forsythia* (reported pathogens) and proposing non-culturable spirochetes and *Filifactor alosis* as biological markers for periodontitis, along with many others (Figure 24). Three such markers (*Fusobacterium*, *Prevotella* and *Selenomonas*) were differentially increased in smokers from an exploratory study by Bizzarro and collaborators, supporting them as causative agents under dysbiosis (*191*).

Borrowing from comparative metagenomic approaches developed for the gut, the group of Alex Mira published the first truly dedicated metagenomic (WGS) study comparing the differences in the healthy versus periodontitis and caries subgingival communities using pyrosequencing of eight samples, producing 1Gbp of sequences that were not biased by amplification (*192*). In this work, Belda-Ferre and collaborators found a functionally distinct composition of the oral microbiota when compared to that found in the gut. Metabolic genes involved in sugar uptake and assimilation, adhesion proteins and prophage genes were more prevalent in the gut whereas the oral metagenome was enriched in gene families for scavenging. More importantly, healthy subjects presented antibacterial peptides like bacteriocins and periplasmic stress response genes. Contrastingly, most prevalent genes in active caries individuals were involved in mixed-acid fermentation and DNA uptake, congruent with the caries formation mechanisms. Individuals that had never presented dental caries were particularly interesting as *S. mutans* was not detected in their plaque microbiota and they had an increased antimicrobial peptide and quorum sensing metagenomic content. The group would then develop this new line of research, isolating and characterizing a novel streptococcus from the mitis group they named *S. dentisani* (*193*). This organism presents an elevated production of bacteriocins and a marked buffer activity under acidic pH, effectively lowering suitability of *S. mutans* and other cariogenic organisms (*194*) as these bacteria decrease the pH of the oral cavity due to their carbohydrate metabolism, resulting in the demineralization of tooth surfaces (*195*).

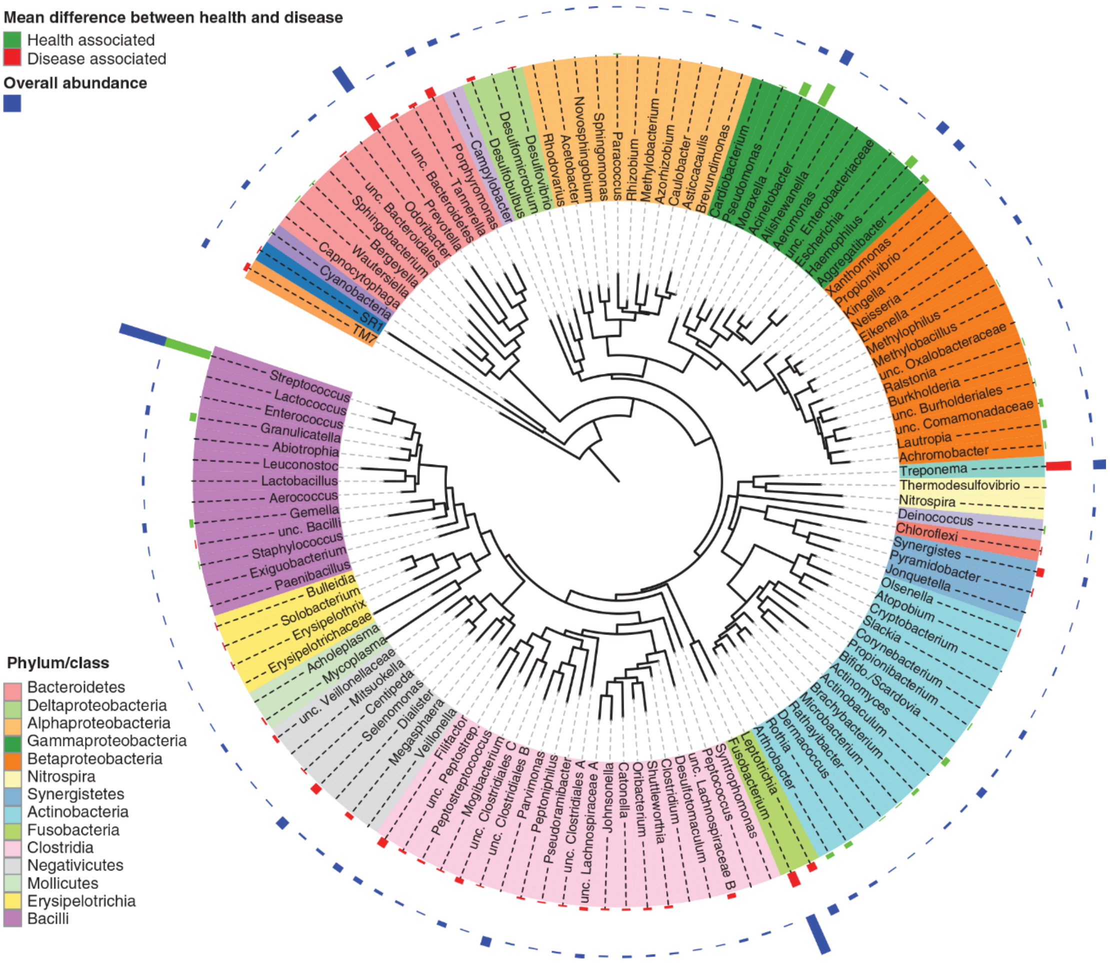


Figure 24. Maximum likelihood phylogenetic tree at genus level coloured by different phylum/class. The health and disease mean differences are shown in the outer circles as well as overall abundance. From: Griffen et al., 2011 (*190*).

Precisely, metagenomics produced a renewed interest in bacterial interactions (with each other and with its host) as it was now clear they were not acting alone and, in fact, the complexity of the different niches was much higher than originally expected. Within a niche, bacteria can communicate by liberating chemical signals to the medium in a process called quorum sensing that is believed to play an important role in the formation of dental plaque (resulting from cooperative aggregation) and modify the production of bacteriocins by streptococci (e.g. The ComCDE system in *S. mutans* regulates bacteriocins by sensing competence stimulating peptides from the medium), some of the main colonizers in dental biofilm (*196*). Regarding the oral microbiota's interaction with its host, it is known that the inflammation response can be modulated by bacteria. Neutrophils play a major role in the gingivae, can establish a gradient in the GCF in response to chemokines such as IL-8 and are recruited to the mucosal tissue by chemokines and cytokines (*197*). Resident bacterial may trigger neutrophil deployment in the subgingival plaque by regulating low levels of expression of intracellular adhesion molecule 1 (ICAM-1), E-selectin, and IL-8 or promote expression of IL-1B mRNA in the oral mucosa. Known mechanisms for oral bacteria to modulate or suppress inflammation include signalling Toll-like receptors or NOD-like receptors, inhibiting activation of NF-κB or secreting IL10 and other anti-inflammatory cytokines. Periodontal neutrophils depend on CXCR2 and its ligand is up-regulated by commensal colonization. Up to 30-40% of the resident streptococci from the tongue or plaque were able to inhibit IL-8 secretion (largely via inhibition of NF-κB) from cells stimulated by flagellin, LL-37 or by oral pathogens such as *P. gingivalis* and *A. actinomycetemcomitans* (*197*). This has also led to consider different strategies to control dysbiotic populations other than antibiotic intake, mainly because it has been recognized that not all bacteria play a harmful role to health (*198*). Thus, in order to keep beneficial consortia, different strategies based on restoring balance may be applied, such as pH modulation and beneficial probiotics (*194*).

A much better sense of the community structure of microbiota can be seen in a recent study by Welch and collaborators (*199*). In it, they analyse the spatial organization of the plaque microbiota using state-of -the-art spectral imaging fluorescence in situ hybridization has provided first-hand evidence of the three-dimensional structure of the oral biofilm. The group has called this a hedgehog structure due the much visible array of *Corynebacterium* filaments (Figure 25a). Different layers, called base, annulus and perimeter are occupied by varying taxa, ordered by oxygen availability (anaerobes in the inside). Consumers and producers of metabolites like lactate tend to be close together in the same layers.

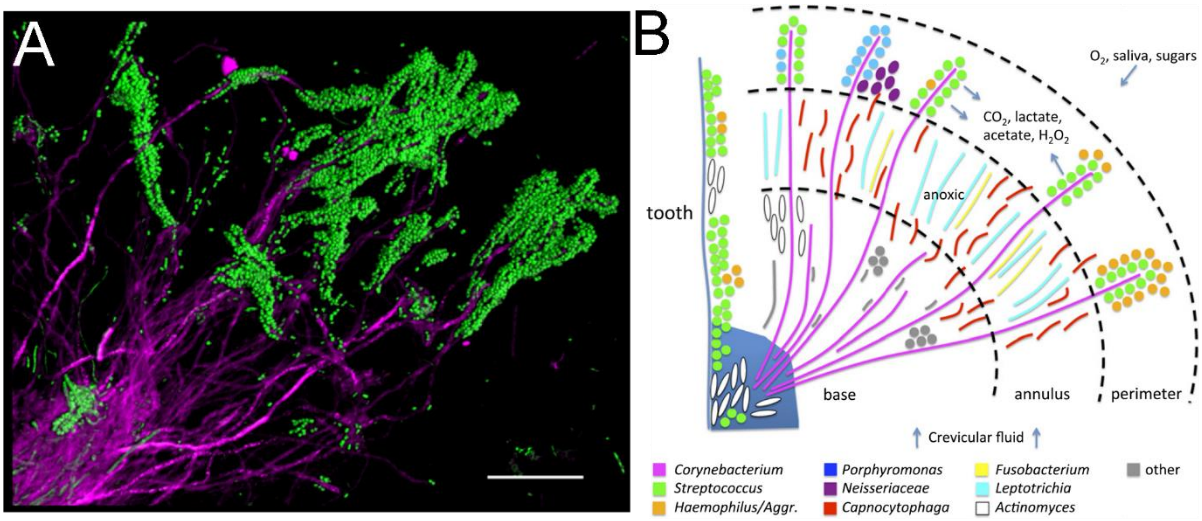


Figure 25. Hedgehog structures of oral plaque. A) Dual probe hybridized structure. *Corynebacterium* cells (magenta) are seen as long filaments. Cocci (green) are bound to the tips. B) Interpretation of the structure. *Corynebacterium* filaments are capped by *Streptococcus* and *Porphyromonas* in the periphery, along with aggregated *Haemophilus*/*Aggregatibacter* and *Neisseriaceae*. *Streptococcus* creates microenvironment rich in CO_2, lactate, and acetate, containing peroxide, and low in oxygen. Elongated filaments of *Fusobacterium* and *Leptotrichia* proliferate inside the outer shell in an annulus. The base is dominated by *Corynebacterium* filaments, also populated by other rods and cocci. Modified from: Mark Welch *et al.*, 2016 (*199*).

As a summary of the different studies comparing diseases, Table 4 includes the most common organisms associated to them.

Table 4. Microbiota of the human mouth in health and disease. From: Williams *et al.*, 2014 (*200*)

Health	Gingivitis (continued)	Chronic periodontitis (continued)
Teeth	*Actinomyces viscosus*	*Porphyromonas endodontalis*
Streptococci	*Streptococcus sanguinis*	*Wolinella recta*
Streptococcus mitis bv. 1	*Fusobacterium nucleatum*	*Treponema* sp. strain 1:G:T21
Streptococcus gordonii	*Selenomonas sputigena*	*Fusobacterium nucleatum*
Veillonellae	*Haemophilus parainfluenzae*	*Atopobium rimae*
Streptococcus sanguinis	*Actinomyces israelii*	*Megasphaera* sp. clone BB166
Streptococcus oralis	*Streptococcus mitis*	*Catonella morbi*
Actinomyces	*Peptostreptococcus*	*Eubacterium saphenum*
Tongue	*Prevotella intermedia*	*Gemella haemolysans*
General streptococci	*Campylobacter sputorum*	*Streptococcus anginosus*
Streptococcus mitis bv. 2	*Veillonella species*	*Campylobacter gracilis*
Streptococcus salivarius	Chronic periodontitis	*Haemophilus parainfluenzae*
Disease	Clone I025	*Prevotella tannerae*
Dental caries	*TM7*	*Porphyromonas gingivalis*
Streptococcus sanguinis	*Fusobacterium nucleatum* subsp. *animalis*	*Peptostreptococcus micros*
Streptococcus oralis	*Atopobium parvulum*	Localized aggressive periodontitis
Mutans streptococci	*Eubacterium* sp. *strain* PUS9.170	*Eikenella corrodens*
Veillonellae	*Abiotrophia adiacens*	*Capnocytophaga sputigena*
Streptococcus mitis bv. 1	*Dialister pneumosintes*	*Aggregatibacter actinomycetemcomitans*
Streptococcus gordonii	*Filifactor alocis*	*Prevotella intermedia*
Actinomyces	*Selenomonas* sp. strain GAA14	
Lactobacilli	*Streptococcus constellatus*	
Gingivitis	*Campylobacter rectus*	
Actinomyces naeslundii	*Tannerella forsythia*	

Oral lesions

Oral leukoplakia and malignant transformation

White plaques ("leuko-plakia") resulting from hyperkeratosis and acanthosis (diffuse epidermal hyperplasia causing thickening of the tissue) in the mucosa of the oral cavity are recognized as potentially precancerous lesions by the World Health Organization (WHO) (*201*). Those that cannot be characterized clinically or pathologically as any other disease presenting similar patches (lichen planus, candidiasis, white sponge nevus, etc.; Table 5) are diagnosed as oral leukoplakia (OL). OL is generally painless and may precede the appearance of cancer by months, or years, or may be present together with the carcinoma. With an estimated prevalence of 2.6% in the worldwide population, the crude annual oral cancer incidence rate due to leukoplakia for the 2000s was between 6.2% and 29.1% depending on the type of lesion (*202*). The prevalence of OL was found to be significantly higher in India (3.28%), probably due to the widespread consumption of paan, a stimulant psychoactive preparation made from areca nut wrapped in betel leaf that is chewed or swallowed, similarly to tobacco (*203*). OL correlates with age, in developed countries it is most common after 40 years of age whereas in developing countries it is diagnosed starting at 30 (*204*).

Table 5. White lesions of the oral cavity (other than leukoplakia) From: Villa and Woo, 2017 (*205*)**.**

Developmental	Cannon white sponge nevus
	Hereditary benign intraepithelial dyskeratosis
	Other congenital genodermatoses (e.g. pachyonychia congenita)
Reactive or frictional	Leukoedema
	Contact desquamation
	Frictional keratosis: MMO, BARK
	Hairy tongue
	Associated with tobacco use: nicotinic stomatitis, smokeless tobacco keratosis
Infectious	Candidiasis
	Hairy leukoplakia (associated with EBV)
Immune mediated	Lichen planus
	Lichenoid lesions
	Benign migratory glossitis
Autoimmune	Lupus erythematosus
	Chronic graft-vs-host disease
Metabolic	Uremic stomatitis
	Palifermin-associated hyperkeratosis
Malignant and OPMD	Keratosis of unknown significance
	Dysplastic leukoplakia
	SCC
	Verrucous carcinoma

Abbreviations: BARK, benign alveolar ridge keratosis; EBV, Epstein-Barr virus; MMO, morsicatio mucosae oris; OPMD, oral potentially malignant disorders; SCC, squamous cell carcinoma.

OL patches were extensively described in a report by the WHO in 1978 about precancerous lesions (*201*). These can be white, slightly yellow or grey (Figure 27). They may appear homogeneous, nodular (wart-like), or specked with white excrescences on an erythematous base (abnormal redness). Tobacco has been pointed by the WHO as an etiological agent for some of these lesions (not to confuse with non-cancerous stomatitis nicotina, a reversible hardening of the palate in smokers) and notes they may regress if smoking or chewing is discontinued. Histopathologically, OL lesions are variable and may present hyperorthokeratosis or hyperparakeratosis, with some degree of diffuse chronic inflammation infiltration in the lamina propria usually containing both lymphocytes and plasma cells. Some lesions may present acanthosis and/or poorly defined borderline limits. Epithelial dysplasia (abnormal growth and cell maturation) may be present in OL lesions, more commonly in the nodular type.

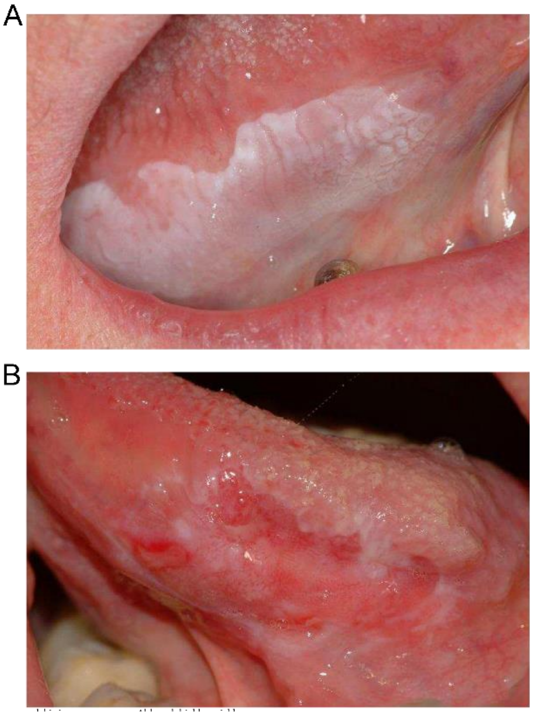


Figure 26. Varying appearance of oral leukoplakia. A) Homogeneous leukoplakia of the lateral border of the tongue. B) Non-homogeneous leukoplakia of the lateral border of the tongue. Adapted from: Lodi *et al.*, 2006 (*206*).

Within lesions, different cellular changes (atypia) may occur in the squamous cell epithelium resulting in dysplasia, often regarded as step preceding malignant transformation (*201*). Those listed by the WHO include the loss of polarity of the basal cells, one or more cell layers having basaloid appearance, increased nuclear-cytoplasmic ratio, drop-shaped rete processes (invaginations of the epithelium into the connective tissue), irregular epithelial stratification, increased mitotic figures, mitotic figures in the superficial half of the epithelium, cellular pleomorphism, nuclear hyperchromatism, enlarged nucleoli, reduction of cellular cohesion, keratinization of single cells or cell groups in the prickle layer. Even so, minor degree atypia is often present in inflammatory conditions and thus a low degree dysplasia in not suggestive of malignancy. However, dysplasia appearing in the floor of the mouth and the ventral surface of the tongue are considered high-risk indicators.

The varying appearance and multiple probable causative factors of leukoplakic lesions makes long-term treatment difficult and prognosis is generally poor (*207*). In an effort to assess the malignancy rate of OL Silverman and collaborators carried out a prospective study on 257 patients whose clinical evolution was followed for eight years, on average, after being diagnosed. They reported 17.5% developed squamous carcinomas, confirming a high risk for malignant transformation of the lesions (*208*). Interestingly, of the 45 patients that developed malignant transformation, some presented a particularly recurrent precancerous subtype with verrucous hyperplasic papillae, described as fast spreading irregular exophytic growth with sharp or blunt warts with most also displaying an erythematous component.

A more aggressive form – the case of proliferative verrucous leukoplakia

A year later, Hansen and collaborators from the Silverman's group, published a more specific retrospective analysis of 30 patients that developed verrucous subtype in their previous study (*207*). These patients showed a continuum of disease ranging from simple hyperkeratoses to invasive oral squamous cell carcinoma (OSCC) or verrucous carcinoma (VC) capable of metastasis. They named the associated lesions *proliferative verrucous leukoplakia* (PVL), a condition described as an aggressive form of oral leukoplakia characterized for its increased malignancy (with a rate of 40-100% depending on the study,) and resilience (Figure 27). As the group noted, PVL is slow-growing and commonly begins as a simple hyperkeratosis in the oral mucosa but tends to spread and become multifocal (appearing simultaneously in different mucosal tissues). It is persistent and irreversible, and may develop warts in the surface epithelium of the affected area. Diagnosis must be made histologically and adequate biopsy specimens are necessary (Box 5). In the early stages of the lesion, PVL is particularly difficult to diagnose as it may not be distinguished from treatable types of OL (Figure 21). Also, biopsies in such early lesions are presumably benign in histological appearance, normally not presenting dysplasia. Contrastingly, neither speckled or nodular leukoplakia present a long history of keratosis without dysplasia.

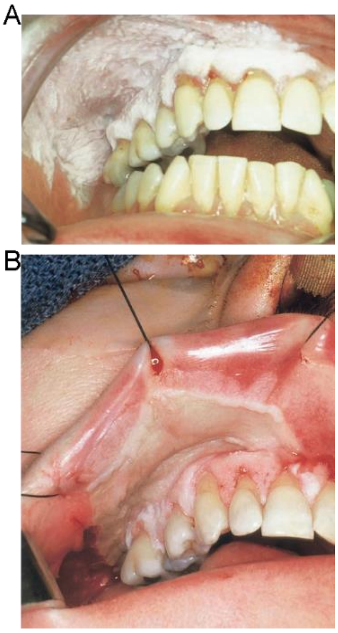


Figure 27. Recurrence of proliferative verrucous leukoplakia. A) Initial presentation of the PVL lesion covering the right anterior and posterior maxillary gingiva and right buccal mucosa extending posteriorly behind the last molar. B) Recurrence of the lesions at the periphery of the skin graft sparing the skin graft itself. Adapted from: Vigilante *et al.*, 2003 (*209*).

Diagnoses cannot be confirmed until multifocality and resilience to treatment are confirmed. The last stages (8-10 in Figure 28) cannot be easily differentiated from OSCC (Figure 29). Malignancy cannot be attributed from the observations of PVL lesions alone as there is no clear physiological pattern for its diagnostic. As Hansen reported in virtually all patients, the least malignant areas that were detected appeared as simpler hyperkeratotic patches of OL (Grade 2 in Figure 28) whereas the most malignant varied, with most having papillary squamous malignant growth (Grade 8).

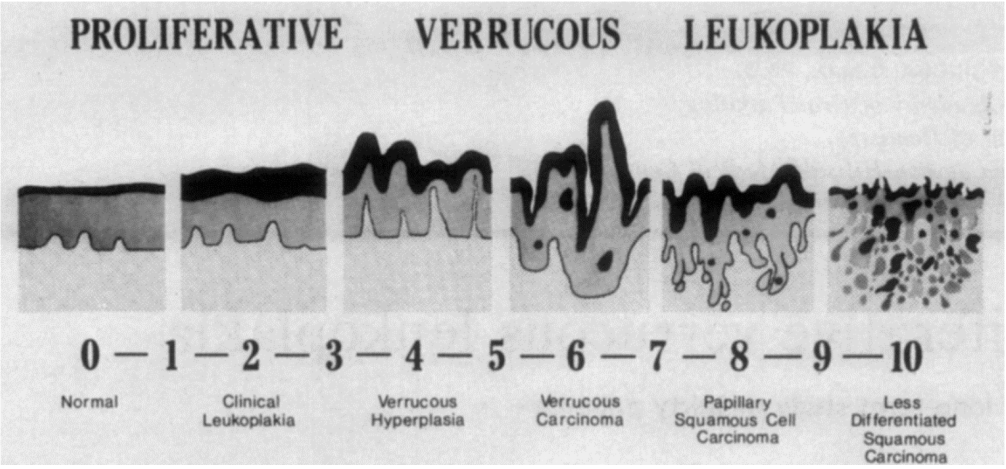


Figure 28. Diagram illustrating the pathohistological grading criteria employed by Hansen and collaborators to classify proliferative verrucous leukoplakia samples in different diagnosis stages. Progression goes form unaffected oral mucosa to undifferentiated tumour cells. Grade 0: Unaffected mucosa. Grade 2: regular hyperkeratosis with little or no dysplasia. Grade 4: papillary exophytic proliferation and little or no dysplasia. Grade 6: slight cancerous appearance, downgrowth of well-differentiated squamous epithelium broad, blunt rete ridges with intact basement membranes, invasion of lamina propria and little or no dysplasia. Grade 8: cancerous appearance, exophytic and invasive growth

and well-differentiated squamous epithelium with keratin formation, narrower invasive fingers of epithelium, less distinct membrane and minimal dysplasia. Grade 10: evident carcinoma, loss of cohesion of moderately or poorly differentiated tumour cells, moderate to severe dysplasia and keratin formation minimal or absent. Indistinguishable from some squamous cell carcinomas. Intermediate grades were used for biopsies that did not fulfil listed criteria. From: Hansen *et al.*, 1985 (*207*).

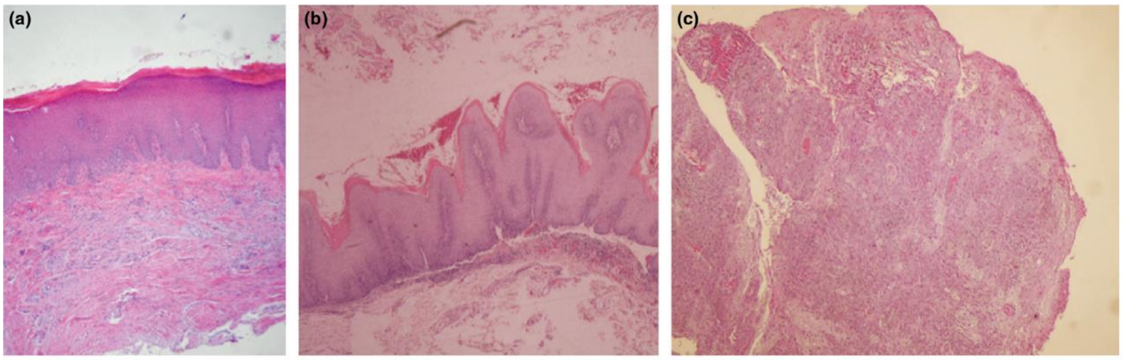


Figure 29. H&E-stained sections of the lesions in groups oral leukoplakia (OL), proliferative verrucous leukoplakia (PVL) and oral squamous cell carcinoma (OSCC). Haematoxylin and eosin stain of histological sections of affected patients. A) Patient of group OL displaying hyperorthokeratosis without dysplastic changes. B) Patient of group PVL with wavy hyperorthokeratosis, and exophytic warty configuration. C) Patient of group OSCC showing conventional squamous cell carcinoma. From: García-López *et al.*, 2014 (*210*).

Box 5 – Diagnosis of PVL.

In 2010, Cerero-Lapiedra and collaborators (*211*) published the following guidelines for diagnosing PVL:

Major Criteria (MC):

A. A leukoplakia lesion with more than two different oral sites, which is most frequently found in the gingiva, alveolar processes and palate.

B. The existence of a verrucous area.

C. That the lesions have spread or engrossed during development of the disease.

D. That there has been a recurrence in a previously treated area.

E. Histopathologically, there can be from simple epithelial hyperkeratosis to verrucous hyperplasia, verrucous carcinoma or oral squamous cell carcinoma, whether in situ or infiltrating.

Minor Criteria (mc):

a. An oral leukoplakia lesion that occupies at least 3 cm when adding all the affected areas.

b. That the patient be female.

c. That the patient (male or female) be a non-smoker.

d. A disease evolution higher than 5 years.

In order to make the diagnosis of PVL, it was suggested that one of the two following combinations of the criteria mentioned before were met.

1. Three major criteria (being E among them) or

2. Two major criteria (being E among them) + two minor criteria.

Hansen and collaborators reported some of the basic features of PVL, same which have been confirmed by similar results in subsequent studies (*212–215*) and are shown in Table 6. The lesions are more common in elderly women (87.5% of the patients were women of over 60 years of age in Hansen's study) and PVL may be a life-long condition (in the study, many patients had had hyperkeratotic growths long before the first biopsy was taken and PVL was still present in survivors at the end of the study). Also, neither tobacco consumption nor *Candida* infections are considered causative agents (although *Candida albicans* was detected in 12 patients). In order of detection, the most common sites presenting PVL lesions include the mucosa of the hard and soft palate, alveolar mucosa, tongue, floor of the mouth, gingiva, and lips but they often occur simultaneously. Aggregated data from 17 studies published until 2014 reported 322 female and 128 male patients with an average age of 63.2 years (*215*).

Table 6. Main differences between localized leukoplakia and proliferative verrucous leukoplakia. Adapted from: Villa and Woo, 2017 (*205*).

Localized Leukoplakia	Proliferative Verrucous Leukoplakia
Mostly in men	Mostly in women
Strong association with cigarette smoking	Weak association with smoking
Single site, usually ventral tongue, floor of mouth	Multifocal
~40% show dysplasia or SCC at time of first biopsy examination	<10% show dysplasia or SCC at time of first biopsy examination, mostly atypical verrucous hyperplasia or KUS
Malignant transformation 3-15% overall	Malignant transformation 40-100% overall
Malignant transformation 1-3% per year	Malignant transformation 10% per year
Easy to ablate or excise because localized	Difficult to treat because multifocal

Most treatment methods have been demonstrated to be ineffective in a long term for controlling PVL (*211*). Some methods that have been used include conventional surgery, laser surgery, radiotherapy, chemotherapy, retinoids, photodynamic therapy, maxillectomy, and stripping (*207*, *212*, *216–219*). Recently, two separate cases of presumably successful treatment of PVL lesions respectively using topical 5-ALA photodynamic therapy (a photochemical light-induced reaction; 12-month follow-up) and topical imiquimod (an immune response modifier; 6-month follow-up) but not conclusive results were reported (*220, 221*).

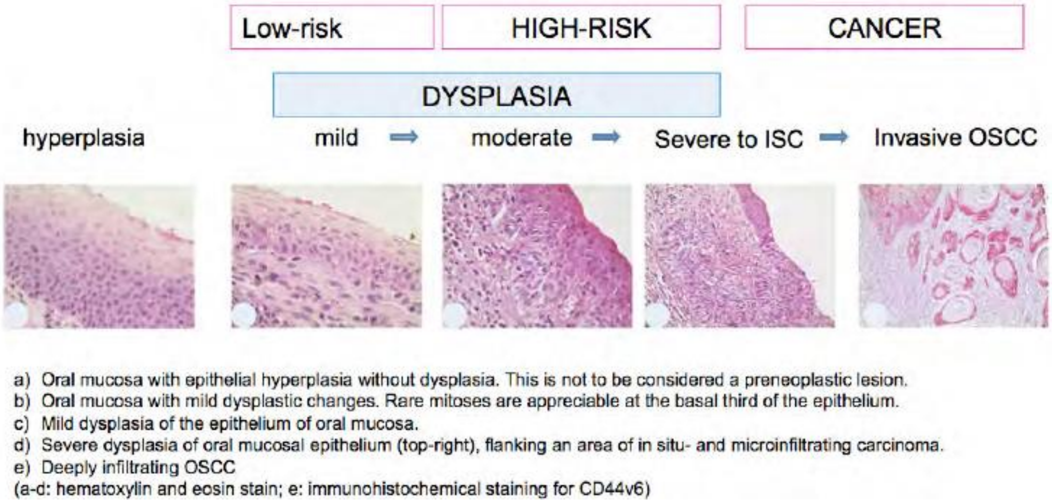
The aetiology of PVL remains ultimately undetermined. In 1994, Roman and Sedano carried out. However, the recurrent and multifocal nature of most PVL lesions resembles multifocal epithelial hyperplasia, or Heck's disease, an infection of the oral mucosa produced by Human Papillomavirus (HPV) types 13 and 32 producing benign papillomatoid lesions in the lips, lining mucosa and tongue (*222*). This, added to their high malignancy and the histological similarity of OSCC and cervical carcinoma (which is incidentally associated to HPV infection) some research groups have pursued the search of a viral agent (possibly an oncovirus, a cancer-causing virus) as possible contributor or causative agent (*223*). The search for HPV in the oral cavity has had variable, and often contradictory results. In 1995, for instance, using PCR amplification as a method of detection, Palefsky and collaborators found HPV-16 in seven out of nine samples and HPV-18 in another, both of which are high-risk types associated with cervical carcinoma (*223*). Some years later, Gopalakrishnan and collaborators identified HPV-16 and HPV-18 in two out of ten subjects (*224*). Fettig and collaborators found only a low risk HPV-11 in one out of ten subjects (*217*) while a more recent study by Bagan and collaborators found none in a 13 patient cohort (*225*). Studies focused in other oncoviruses include one by Bagan and collaborators in which they detected Human Herpesvirus 4, also known as Epstein-Bar virus (EBV or HHV-4), a virus that has been associated to hairy leukoplakia (an unrelated corrugated type) (*226*), via nested PCR amplification in six out of ten patients bearing PVL lesions who had developed OSCC but could not define it as a causative agent as it was also found in two of five patients with OSCC but not PVL (*227*). New approaches, such as metagenomics, may help elucidate an infectious agent by exploring a broader spectrum of viruses and possibly other potential pathogens such as bacteria or fungi.

Oral squamous cell carcinoma

OSCC comprises 95% of all head and neck carcinomas. With over 200,000 cases each year (*228*), 90% of these are associated to alcohol and tobacco consumption (*229*). However, the majority of OSCC are diagnosed at phase III or IV, after they have grown deeply into nearby tissue and they may have spread to the lymph nodes and potentially to other parts of the body (metastasis) (*230*), reducing the chance of survival to an average of 55% (*231*), and this is without considering that a high percentage of patients have a poor response and high recurrence rates (*232*). This has prompted the medical and scientific community to research methods for early detection and prevention. Currently, its most frequently applied clinical diagnoses are oral leukoplakia and erythroplakia (*233*).

OSCC is a malignant neoplasm that derives from the stratified squamous epithelium of the oral mucosa. As in other types of carcinoma, cells presenting atypia acquire some of the following capabilities: insensivity to anti-grow signals, self-sufficiency in growth signals, evading apoptosis, sustained angiogenesis, limitless replicative potential, tissue invasion and metastasis (*234*). Thanks to the study of the cervical neoplasm, it is now known that carcinoma arises from precursor lesions in a long continuum of progressively more atypical changes. The lesion passes through various phases before it establishes as invasive carcinoma *in situ*, which are jointly addressed as preneoplastic damage, showing mild, moderate or severe dysplasia (depending on the thickness of the squamous epithelium presenting atypia), as well as epithelial changes that may include hyperkeratosis, hyperplasia and acanthosis (Figure 30)(*233*). The earliest change is the appearance of atypical cells in the basal layers of the squamous epithelium alongside normal differentiation toward the prickle and keratinizing cell layers and eventually ends with most cells presenting atypia and a non-differentiated surface.

The cancer originates from dysplasia of the squamous cells on the surface of the epithelial layer and subsequently invading the subepithelial basement membrane through islets and cords of epithelial cells, resulting in local destruction and invasion via metastasis.



a) Oral mucosa with epithelial hyperplasia without dysplasia. This is not to be considered a preneoplastic lesion.
b) Oral mucosa with mild dysplastic changes. Rare mitoses are appreciable at the basal third of the epithelium.
c) Mild dysplasia of the epithelium of oral mucosa.
d) Severe dysplasia of oral mucosal epithelium (top-right), flanking an area of in situ- and microinfiltrating carcinoma.
e) Deeply infiltrating OSCC
(a-d: hematoxylin and eosin stain; e: immunohistochemical staining for CD44v6)

Figure 30. Progression of oral squamous cell carcinoma (OSCC). ISC: Carcinoma *in situ* showing thick dysplasia. From: Celetti *et al*., 2012 (*233*).

Genetic changes and sequence of genetic events guiding the progression of normal mucosa to oral neoplastic tissue are not entirely understood but it is estimated six to ten independent events within a single cell lead to OSCC. The process of tumorigenesis involves the inactivation of tumour suppressor genes and activation of proto-oncogene products (*235*). There are several chromosomal alterations in the progression of oral intraepithelial neoplasm. Chromosomal loss was detected to increase progressively at each histopathological step from benign hyperplasia to carcinoma *in situ* and later OSCC. Earliest alterations appear on chromosome 9p21 where gene p16 is found, at 3p that contains three putative tumour-suppressor loci, and at 17p13 where the p53 gene is located. Some other genes that are critically altered in OSCC include cyclin D1, retinoblastoma, epidermal growth factor receptor, signal transducer and activator of transcription 3, and vascular endothelial growth factor receiver (*236*). Over expression of HOXB7 gene, a critical regulator of development, has also been associated to neoplasms in head and neck and is associated to poor prognosis in OSCC (*237*). A more commonly recognized markers for cancer in general is the gene for the p53 protein, a transcription factor that is implicated in cell cycle control (in response to chromosomal damage, p53 turns off the cell cycle and stimulates expression of DNA repair proteins), apoptosis and preservation of genetic stability (more than 50% of OSCC cases have mutated p53 gene) (*238*). Its mutation (occurring in over 50% OSCC cases) prevents the accumulation of cells with DNA damage and is also involved in hypoxia and oncogene activation as its product protects against tumour formation. A marker for recurrence is mutated Ki-67, a gene coding for a protein involve in cell division, particularly in the M phase (*239*). Mutations in ColIV, coding for collagen type IV are key to infiltration of cancer cells and metastasis. This compromises the molecular assembly of the subepithelial basement membrane as collagen type IV is its most important structural component, resulting in its the degradation between the epithelium and the lamina propria (*240*). Furthermore, the matrix metalloproteinases (MMP)-2 and MMP-9 degrade ColIV and mutations in their genes have also been reported to in cases of OSCC.

HPV infection has been recognized as an increased risk of OSCC, with the most common types being HPV-16 (90% of all detected cases) and HPV-18 and low risk HPVs associated with benign warts (*241*). In a review, Zaravinos pointed out the detection rate is between 20-40% in most studies. The mechanism by which HPV infection may contribute to OSCC is not clear yet but it has been hypothesized that HPV may contribute to the accumulation of chromosomal mutations. The E6 and E7 proteins of HPV have been shown to bind and inactivate p53 and retinoblastoma proteins respectively (*223*).

In its most recent estimates (in 2012), the International Agency for Research on Cancer reported that around 15.4% (2.2 million) of all new cancers detected worldwide each year are attributable to carcinogenic infections (16.1% of all current), most notably by *Helycobacter pylori* (770,000), high-risk human papillomaviruses (HPV; 640,000), hepatitis B virus (HBV; 420,000), hepatitis C (HCV; 170,000)and Epstein-Bar (EBV; 120,000) while other are more localized like Kaposi's sarcoma as the second largest in Africa and southern Europe (*228*). Of the total 200,000 new cases of oral carcinoma, the fraction attributable to oncoviruses is estimated at 4.3% but only HPV, mainly types 16 and 18 have been attributed as a risk in the oral cavity which are commonly detected by the PCR amplification of genes for the E6/E7 proteins. The agency recognizes eleven infectious agents (including oncoviruses) as well established carcinogenic agents in humans: *H. pylori*, HBV, HCV, Human immunodeficiency virus 1 (HIV-1, only in presence of other carcinogens), several human papillomaviruses (HPV types 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, and 59) collectively known as high-risk types, two herpesvirus: EBV (HHV-4) and Kaposi's sarcoma (human herpesvirus type 8, HHV-8), human T-cell lymphotropic virus type 1 (HTLV-1), and eukaryotic parasites *Opisthorchis viverrini*, *Clonorchis sinensis*, and *Schistosoma haematobium*. The more recently discovered Merkel cell polyomavirus (MCV or MCPyV, human polyomavirus 5) has also been implicated in carcinomas of skin and mucosa.

Recent years have seen the first metagenomic studies for the identification of infectious agents in OSCC but most have been focused on bacteria. The first of its kind was a pyrosequencing analysis by Pushalkar and collaborators in 2011 (*242*). They studied the diversity of bacteria in the progression of OSCC in the floor of the mouth of three smokers and two healthy controls using a 16S profiling approach. In accordance to previous PCR assays, they found high levels of colonization of OSCC by facultative oral streptococci, plus they detected 15 allegedly unique phylotypes, including pathogenic bacteria *Capnocytophaga* spp., *Fusobacterium necrophorum*, *Prevotella melaninogenica*, *Porphyromonas gingivalis*, *Prevotella intermedia*/*nigrescens*. However, due to the small sample size, their results were not determinant. The same applies for a subsequent study by the same group in which they published a comparison of tumour and non-tumour tissues in 5 patients with OSCC. They found a higher prevalence of streptococci in tumour sites with respect to controls that were taken from upper aerodigestive tract (*243*). In 2014, Roberson and collaborators designed a ~60,000 probe assay called PathoChip derived from partial sequences from pathogenic viruses, bacteria and fungi for scanning carcinomas, for culture free wide-spectrum experiments and studied OSCC. They reported the successful identification of HPV-16i in OSCC samples with metagenomic extractions (*244*). A larger study by Noor Al-hebshi in 2017 analysed 20 OSCC biopsies and 20 controls by 16S profiles carried out with an Illumina platform. They detected higher prevalence of *Fusobacterium nucleatum* and *Pseudomonas aeruginosa* in OSCC and lower numbers of genera *Streptococcus* and *Rothia* (*245*).

Viruses

Beyond the veil

In spite of their ubiquity and the relevant ecological role they play in almost any ecosystem, including the human body, the study of viruses as a branch of metagenomics (viral metagenomics or viromics) has been notably lagging behind prokaryote-focused studies (Figure 31), most likely because of the technical limitations it must face, its variability and, in general, the lack of development in a field that has still much to explore.

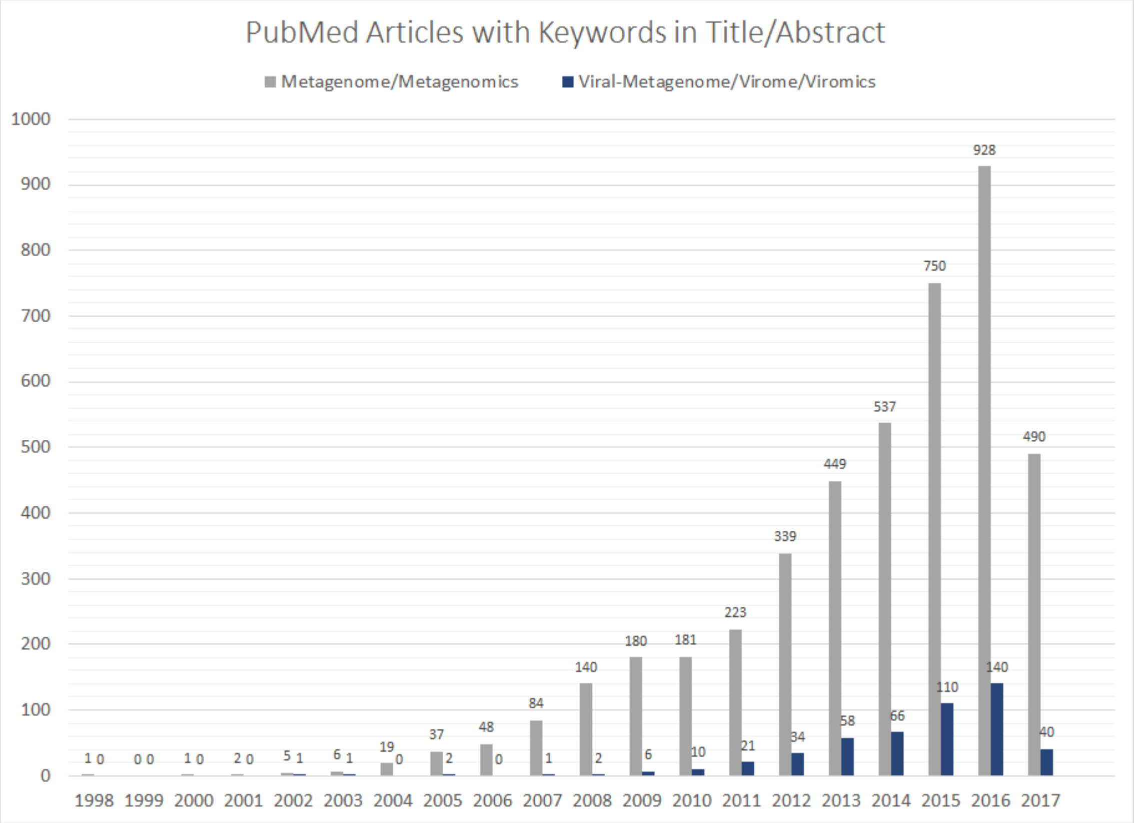


Figure 31. Articles in PubMed bearing the specified keywords in the title or abstract. Data obtained from PubMed on May 14th, 2017.

As stated before, the human body has some $3.8{\cdot}10^{13}$ bacterial cells in a reference male subject (*78*). However, it is believed there may be approximately ten particles for every bacterial cell in aquatic ecosystems, most of them predating bacteria (bacteriophages) (*4*). Whether or not this can be extrapolated to the human body is still unknown as viromic studies face different limitations. Yet, in 2013, Popgeorgiev and collaborators published the results of an analysis of donated blood from ten asymptomatic subjects via WGS and pyrosequencing (*246*). By analysing the viral-enriched fraction they determined the existence of the viral family *Anelloviridae*, comprised of relatively small eukaryotic viruses around 30-50 nm long (*247*), in all samples including *Torque Teno Virus* (TTV), *TTV-like virus*, *SEN virus*, *TTV midi virus*, and *TTV-like mini virus* (*246*). On the other part of the scale, Popgeorgiev managed to recover enough sequences to assemble the genome of a large virus which they suitably called *Giant Blood Marseillevirus*, as it proved related to 250 nm marseillevirus that had been detected to infect amoebas. A significant fraction of the sequences could not be identified due to the lack of homologous references, which is very common in viromic analyses. In fact, that is the reason why some TTV sequences had been previously discarded as PCR artefacts in studies of blood transfusions (*247*). Nonetheless, it is evident the population of human-associated viruses is significantly diverse.

A more recent analysis of the blood virome by Moustafa and collaborators using metagenomic sequences of over 8,200 presumably healthy humans using a greater sequencing depth (more bp generated per subject) found 94 different viruses, including herpesviruses, anelloviruses, papillomaviruses, three polyomaviruses, adenovirus, HIV, HTLV1, HBV, HCV, parvovirus B19, and influenza virus, with 19 of these in 42% of all subjects (*248*). Interestingly, they reported a high prevalence of anelloviruses regardless of geographical background, HTLV1 mostly in subjects from African origin, HHV6A in Middle Eastern subjects (slightly increased in Europeans as well) and MCP in Asian, European, and American subjects. They manage to assembly several of them at least partially. The authors hypothesized the detection of RNA viruses was due to their integration into the host's genome.

In fact, the relationship of the human host with its viruses goes way deeper since they have been partners in genomic evolution, a feature that traces back to the DNA composition of all mammals, and even before the divergence of the group (*249*). Mammalian genomes contain an average of 40% products of reverse transcription, part or which resemble past provirus integration (exogenous retrovirus infection can result in the integration of the viral genome into the host's genome). In fact, many of these insertions are estimated to have occurred some 40 to 90 million years ago (*250*) but in 2017, a controversial study by Aiewsakun and Katzourakis reported the analysis of 36 lineages of basal amphibian and fish foamy like endogenous retroviruses, noting retroviruses may have existed for more than 450 million years, at least since the Ordovician period in the Paleozoic Era (*251*). Characteristically flanked by long terminal repeats, these retro-transcribed viral sequences known as endogenous retroviruses (ERVs) are estimated to make up 8 to 10% (~400,000 loci) of all human and mouse DNA. Some of these are reversible, something that has been extensively studied since in classical experiments such as the reactivation of murine leukaemia virus in AKR mouse strains by Lowy and collaborators back in 1971 (*252*). Their pathogenic role has captivated most of their interest, mainly due to the mutagenic global effect of retrotranscribing mobile elements that may cause (e.g. gene disruption) (*253*). However, the present-day fraction has managed to survive negative selection and loss by genetic drift and has gone to fixation, given by their relationship to their host, suggesting a beneficial role for the host, something that has been found in cases examples of ERV-protein domestications to serve host functions, as demonstrated in placental development in murine models (*254*).

These studies demonstrate some of the vast and highly variable collection of viruses and their relevance to humans. Thus, viruses undeniably form part of the associated microbiome found in humans, comprising a fraction that has for the most part, been put aside until recently. But before diving deeper into the world of viromics, it is important to know what they are, how they thrive and why may they be biologically relevant.

Viruses – A story of world domination

The definition of *virus* has adapted to the changes in a field that has seen some of the most important developments in molecular biology in the since the past century. A virus is currently defined as an infectious, obligate intracellular parasite containing a DNA or RNA genome which can direct the synthesis of viral components by the cellular systems within an appropriate host cell. From these, new independent infectious particles (virions) are self-assembled, liberating (*250*). A progeny virion is the vehicle for transmission to the next host cell or organisms, where its disassembly initiates a new infectious cycle. Viruses are regarded as non-living entities having an inanimate phase, the virion, and a live-like multiplying phase.

In order to put these ideas into perspective, ecological investigations have deemed virus as ubiquitous entities, infecting cellular organisms of all domains of life in every type niche and location, making them the most abundant entities in our planet, with more than 10^{31} particles in total, most of them predating marine bacteria (bacteriophages), a conservative estimate elaborated from some figures by and elaborating on estimates by Whitman and Wommack (*3, 4*). Though individual virions are in fact in the nm range, Rohwer calculated that if stacked side by side (assuming an average 40 nm capsid diameter) the column would stand 42,300,000 light years tall. Their very ample variety of hosts is as varied as their array of molecular mechanisms that they use to highjack a cell's replication machinery, which makes them impressively successful. In fact, Hendrix calculated that over 10^{24} productive viral infections occur per second on Earth (*255*) and killing 20-40% of all marine microbes daily (*250*), releasing their content and nutrients as well as CO_2. This naturally impacts both the global ecology and the evolution of all organisms as 10^{28} bp of DNA are estimated to be transduced per year by marine phages alone (*256*), part of which effectively alters the composition of their host genomes. Together, Viral DNA and RNA comprise 94% of all nucleic acid content in the oceans (*250*). Their own repertoire, however, is somewhat more limited than previously expected as it is probably shared among the many species, based on revised figures in a meta-analysis by Ignacio-Espinoza and collaborators in which they positioned the total global virome to have less than 3.9 million different protein clusters (*257*).

History and discovery of viruses

The following part of this section contains summaries from the book Principles of *Virology 4th Edition* (*250*) unless otherwise indicated.

Due to their particular propagation mechanisms, viruses have also played a critical role in the development of many keystone discoveries and basic techniques in the fields of molecular biology and biotechnology. Recombination, mutation, DNA repair, gene transduction, DNA's role as hereditary unit, restriction and editing techniques, expression, activators, repressors, mRNA, triplet coding, and several other discoveries were made in bacteriophages. Also, studies in animal viruses have shed a light on many of the fundamental principles of cellular function. Biotechnologically, viruses are important vectors and their transduction mechanisms have been amply exploited, and, more recently, the related CRISPR editing tools have been derived from a bacterial defence mechanism against phages.

The fact that endogenous viruses are shared among mammals suggests human-associated viruses have existed since the species themselves appeared. Presumably, two key events occurring around 10,000 years ago marked this co-evolution: animal domestication, providing ground for animal-to-human transmission whenever they crossed the species barrier, and sedentary lifestyle, providing larger populations for virus to experiment change in. Later was commerce, providing new human-to-human transmission. It is believed less virulent viruses like modern retroviruses, herpesviruses and papillomaviruses were first to adapt to small populations, while those with high virulence such as measles or smallpox would spread only after larger populations had been formed, providing higher chance of survivors. Early examples of recorded disease include stone tablets in ancient Egypt depicting poliomyelitis from the 15th B.C., and Mesopotamian laws referring to rabid dogs in the 10th century B.C. Smallpox was probably endemic in the Ganges valley and spread to Europe and other parts of Asia and has resulted in lethal epidemics European wars and the colonization of the Americas. Other viral diseases that were reported throughout early history include influenza, mumps and yellow fever described since the first African explorations.

First attempts at immunization were registered the 11th century in China. They were not aware smallpox was caused by an infectious agent but they noted some people survived and never got infected again after having the disease. As a consequence, they started using variolation, which consists on inoculating people with the pustules of diseased patients in the hopes of them acquiring immunity. The usage of a wild type smallpox virus resulted in a high mortality but the practice spread to the Middle East nonetheless. The practice did not become available in the west world until the 1971 when Lady Mary Wortley Montagu, wife of the British ambassador in the Ottoman Empire, brought this knowledge to England. George Washington is said to introduce variolation to soldiers in 1776. Later, in the 1790s, Edward Jenner managed to immunize children to smallpox in England by exposing them to cowpox, after detecting this related infection was present in the hands of milk maids that never developed smallpox. This establishing the principles of vaccination (the term derived from Latin *vacca*). However, viruses were still unknown and they draw their conclusions from empiric physiological examination of the patients. Despite Leeuwenhoek had already set the idea of microorganisms existing everywhere, and Pasteur demonstrated they replicated on their own when conditions were adequate, it was not until Koch linked them to disease that they were first treated as pathogenic agents, following the germ theory.

However, viruses would remain unknown until the end of the 19th century. The first one to be observed was the tobacco mosaic virus (TMV). Tobacco was a commercially relevant product at that time and there was an important disease affecting European plantations, marking the leaves with whitish pigmentation that render crops unsellable. Scientist had grinded the infected leaves in hopes of finding culprit bacteria but found nothing. Later, at the end of the century and in separate experiments, Ivanovsky and Beijerinck demonstrated the filtered liquid obtained from the grinding could still cause the leaves to be infected after removing bacteria. This contagious or poisonous fluid became known as *contagium vivum fluidum* or later virus from Latin for “poison”.

Later, in 1898, Loeffler and Frosch studied foot and mouth disease viruses, the first animal viruses to be isolated, also obtained by filtration. They managed to determine that not only were these agents very small (passing 0.2µm filters) but that they could only replicate in broth, not isolated. Viral research developed during the 19th century with milestones such as the first human virus to be studied, the yellow fever virus in 1901, the rabies virus in 1903, variola virus in 1906, the detection of the first cancer-inducing virus (oncovirus), chicken leukaemia virus which is a poliovirus in 1908, Rous sarcoma virus in 1911, the discovery of bacteriophages in 1915 and the influenza virus in 1933.

With the invention of electronic microscopy in the 1930s, it was clarified that viruses were not liquid but rather particulate agents that infected cells and had very different morphologies. Early taxonomic classification of viruses did in fact follow morphologic features such as the presence of a lipid membrane (the envelope), the size of the virion and the capsid, or host association. Nowadays, taxonomy has shifted to a phylogenetic approach that relies on exact DNA or RNA sequences and is less fallible. However, some lineages still retain older classifications.

Crop relevant viruses were the first to be studied in detail, in part because they could easily transmit between specimens. It took several tries before scientists could infect animals with human viruses, something that was achieved in 1930 by Theiler. With intermediate host organisms, they could be reproduced in sufficient quantities to analyse some of their physical and chemical features.

Replication

However, most early crucial advancements came from bacterial cultures. In the late 1930, bacteriophages had not been demonstrated to be autonomous but were rather thought as metabolic products of bacteria. The cycle of phage infection and reproduction was extensively studied in *E. coli* (Figure 32), eventually describing the mechanisms of recognition, infection, the translation of their mRNA, the assembly and burst (lysing the host cell to release the progeny virions), a cycle addressed as lytic cycle. The attachment of a virus to a cellular membrane is particularly important as it is mediated through the interaction between the surface of the virus and a molecule present on the surface of the cell. Receptors are specific for each type of interaction. A particle may enter the cell via endocytosis, translocation, or fusion of those having a membrane to the cell's own.

Hershey and Chase demonstrated the nucleic acid was the only factor for viral heredity, discarding that proteins were also required. Since the 1920s it had been noted that bacteriophages did not always kill their hosts, but rather activated lysis spontaneously after a period of bacterial growth. These strains were called lysogenic and lysogeny was defined as an alternative cycle in which viruses integrate their genome into the host’s own. The provirus (in this case a prophage), as it is known when integrated, remains in a latent state, while its host divides normally. Thus, the prophage is inherited vertically during each division. Eventually, the phage may revert to the lytic cycle and reproduce. Subsequent experiments demonstrated the mechanisms of viruses found in animal and plants were fundamentally similar.

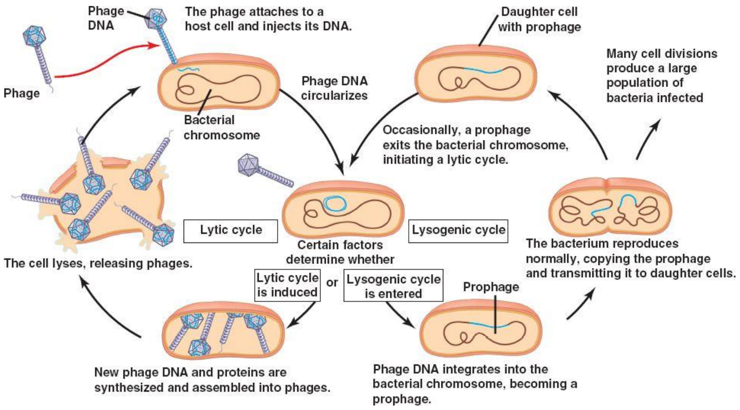

Phage DNA
The phage attaches to a host cell and injects its DNA.
Phage
Bacterial chromosome
Phage DNA circularizes
Daughter cell with prophage
Many cell divisions produce a large population of bacteria infected
Occasionally, a prophage exits the bacterial chromosome, initiating a lytic cycle.
Lytic cycle
Lysogenic cycle
Certain factors determine whether
Lytic cycle is induced
or
Lysogenic cycle is entered
The cell lyses, releasing phages.
The bacterium reproduces normally, copying the prophage and transmitting it to daughter cells.
Prophage
New phage DNA and proteins are synthesized and assembled into phages.
Phage DNA integrates into the bacterial chromosome, becoming a prophage.

Figure 32. Lytic and lysogenic cycle of bacteriophages. Both cycles start with the phage transducing its genetic material into the host cell. Lytic cycle is displayed on the left: Once inside, the genome may be translated into new virion parts, followed by assembly and releasing of new phages. Lysogenic cycle is displayed on the right: the phage genome is integrated into the bacterial chromosome becoming a prophage. This latent state may persist during several cell divisions, cloning the integrated genome as well. Once division ends, the prophage may initiate a lytic cycle. From: http://bio3400.nicerweb.net/bio1151/Locked/media/ch18/18_07LamdaLyticLysoCycle.jpg

Viruses replicate very differently from bacteria (Figure 33). The growth curves of viruses are less like a continuous function as they have to make the parts and then assemble the final product. Just after viral infection occurs, the levels of infectious agents are too low to be detected. This is because during this latent phase, called eclipse period, the genome is liberated from the particle and it is no longer detected as infectious. The genome then replicates and the proteins necessary for its replication and encapsidation are produced. After they are assembled, the cell bursts, expelling a large number of infectious particles, instantly elevating their counts.

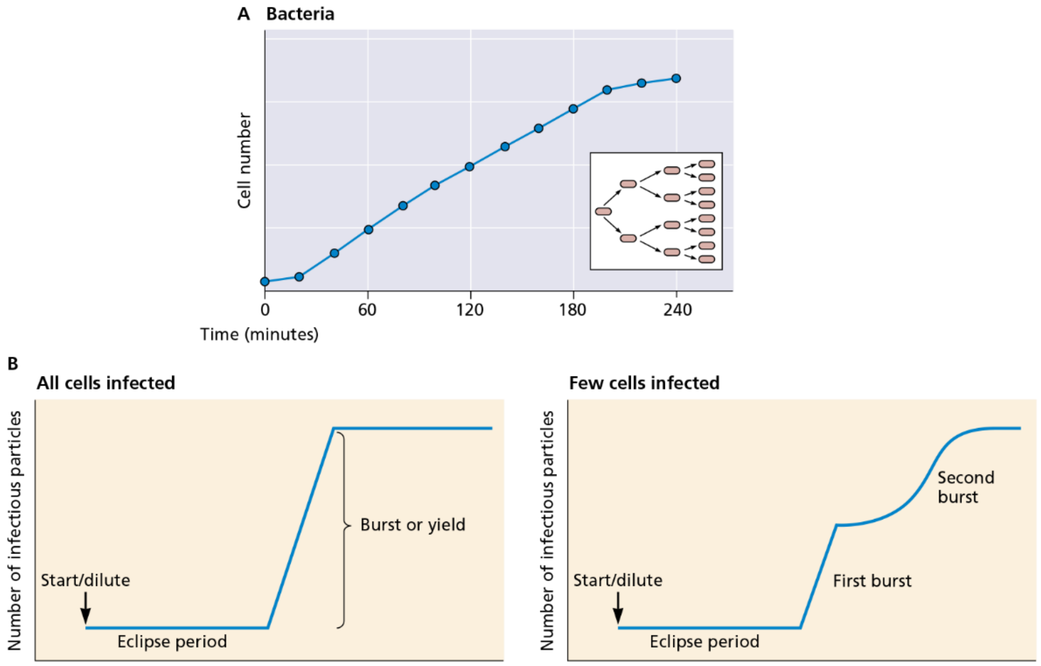


Figure 33. Comparison of bacterial and viral reproduction. The number is plotted as a function of time. A) Growth curve for a bacterium. Starting at one bacterium at time 0, the number of bacteria doubles every 20 min until nutrients are depleted and the growth rate decreases. B) One- and two-step growth

curves of bacteriophages. When all cells are infected at the start of the experiment (left) there is an eclipse period then there is an instantaneous yield or burst of newly assembled virions. When just a few cells are infected (right) a first burst occurs, after which the curve continuously increases in a gradual second burst because the new generation of infectious viruses starts infection at different times. From: Principles of virology 4th Ed. (*250*)

Classification

Viral genomes vary in size, with most ranging from around 2,000 to around 200,000 bases. It may be composed of either RNA or DNA and can be linear or circular or consist of several chromosomes. In the 70s, David Baltimore created a scheme for classifying viruses according to the type of nucleic acid they contain (Figure 34). The central idea behind this system is that known viruses require the translating machinery of a cell to produce its proteins. Thus, they must first produce an mRNA molecule. Of all DNA viruses, only dsDNA (type I) viruses may produce mRNA directly, whereas ssDNA (type II) viruses must first create a complementary (-) strand (type II). Negative strand RNA is the only RNA molecule that can be translated to mRNA directly. RNA viruses with dsRNA (type III) can do this, effectively creating a copy of their + strand, same as the -RNA viruses (type V). On the other hand, viruses with +RNA viruses (type IV) can have their strand directly translated or undergo an intermediate translation first. Others have +RNA that is not directly accessible to ribosomes and must be first retrotranscribed into DNA and get a complementary DNA strand (type VI) before producing an mRNA molecule. *Hepadnaviridae* have a different type of nucleic acid that was missing from the original Baltimore classification, a dsDNA with gaps in one of its strands that must undergo repair before they can produce an mRNA (type VII).

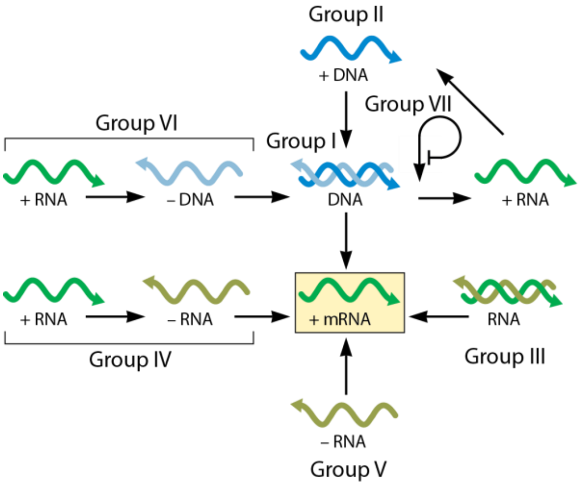


Figure 34. The Baltimore classification. All viruses must produce mRNA that can be translated by cellular ribosomes. In this classification system, the unique pathways from various viral genomes to mRNA define specific virus classes on the basis of the nature and polarity of their genomes. Modified from: Principles of virology 4th Ed. (*250*).

In order to standardize the taxonomy, the International Committee on Taxonomy of Viruses (ICTV) was established as a regulatory authority created in 1966. It was meant to organize viruses based on a set of descriptors that could be applied for assigning taxonomy. At first, classification was based on observation of specific traits of features but names were based on their hosts and even on their discoverers. However, rules for nomenclature did not fully standardized until decades later when sequencing became the norm. Their latest report (2016 v 1.2) recognized the existence of a total of 4,404 species, 736 genera, 36 subfamilies, 123 families, and 9 orders. Yet, most species are probably still missing from the classification system and the collection keeps expanding with each new report (*258*).

Viral taxonomy is broadly based on that of living organisms but has fewer classification levels, ideally starting at the order level (suffix -*virales*), and potentially including family (-*viridae*), genus (-*virus*) and species (highly variable) levels as well as secondary subfamilies (-*nae*) and subspecies. In contrast, serotypes or serovars are classifications based on antigens, a viral phenotypic feature that may be used to classify subspecies.

The scientific community has emitted serious critiques to current virus classification (*259–262*), mainly because taxonomy in viruses poses some serious challenges that have yet to be addressed. There is still a large proportion of uncharacterized particles that have incomplete taxonomic tags, particularly at the levels Order, Family, Subfamily but also at the Genus level, and there is cleaning to do, as several species probably belong to matching genera or other higher levels of taxonomy. As some of these authors note, part of the problem is the slow progress towards accepting new taxonomic categories (the committee meets annually) and that taxonomic information based on viral genetic information alone is complicated to fit into existing taxa due to high rates of horizontal gene transfer (HGT).

Viruses come in very different shapes and sizes (Figure 35). The discovery of giant viruses challenged the previous boundaries of viruses' knowledge. Viruses such as mimiviruses, pythoviruses and pandoraviruses (1.5 micrometres long, which can be seen in light microscopes) are various orders of magnitude bigger than most known viruses but they are still dependent on the replication mechanisms of cells. Regardless, all virus particles have a nucleic acid that is protected by a protein coat, a barrier that protects it from proteases and nucleases, pH or temperature extremes, etc. Protein-protein interactions maintain stable and geometrical (rod-like viruses are built with helical symmetry and spherical viruses with icosahedral symmetry) capsids, some of them comprised of only one type of protein or an aggregate in repeated patterns of more (the more the subunits, the larger the capsid). Rod-shaped viruses' nucleocapsids which integrate nucleic acid and proteins into a regular assembly. Some viruses additionally possess an envelope, which is normally derived from cellular membranes from which they are derived (lipid layer and possibly a protein and glycoprotein layer).

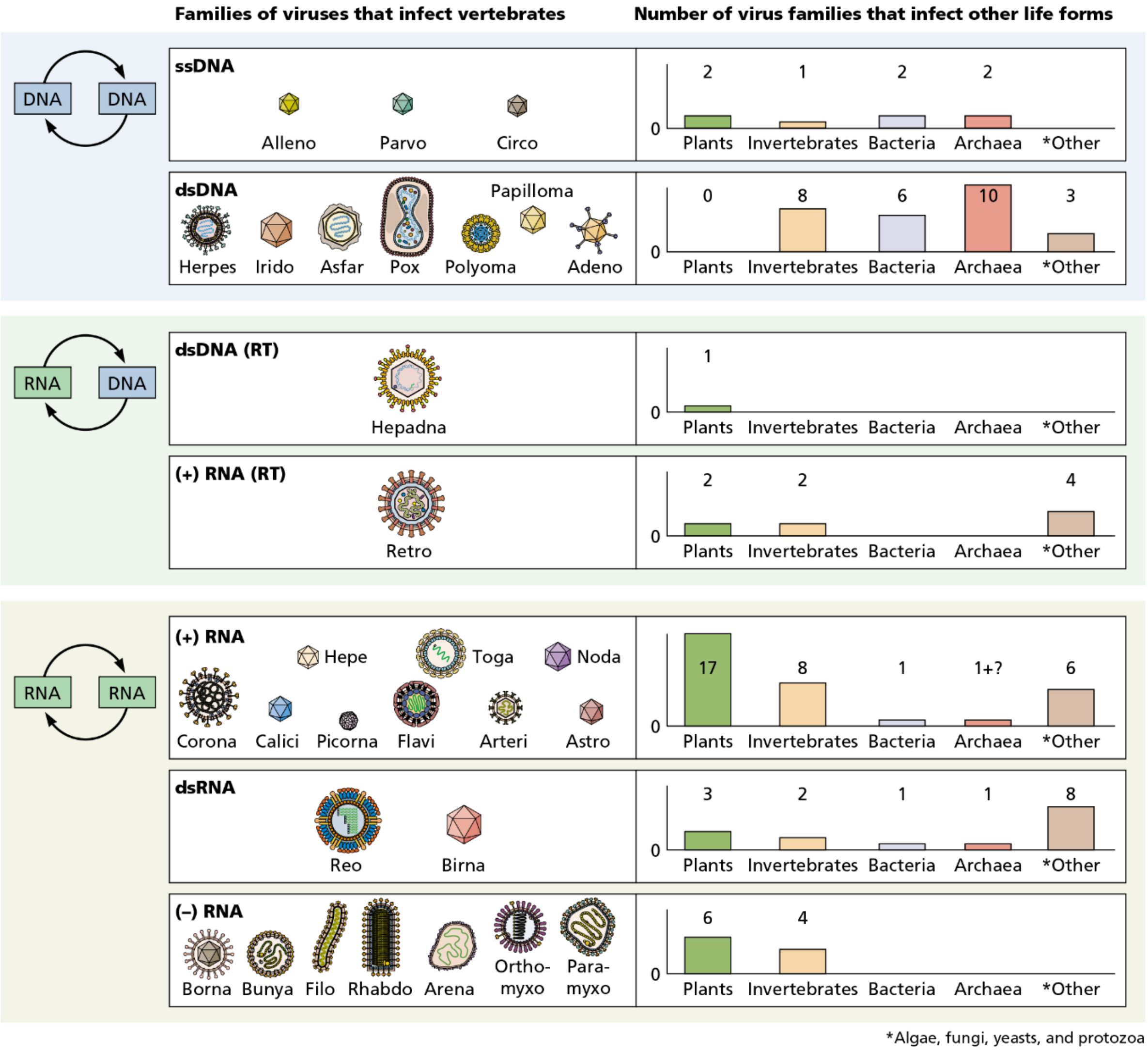


Figure 35. Virus forms and hosts. Viral families are sorted according to the nature of the viral genome. Families infecting vertebrates have been illustrated. In the graphic, the “1+?” notation for RNA archaeal viruses indicates metagenomic analysis only. From: Principles of virology 4th Ed. (*250*).

In general, larger viral particles housing large DNA genomes are structurally more complex presenting different symmetries or multiple layers. The structure of bacteriophages has been studied for 50 years in T4 (Figure 36), a phage infecting E. coli. Its 170 Kbp dsDNA is protected by an icosahedral head built from a single protein. The head is joined to the tail via a connector. The structure derives from a portal, a protein that acts a nanomachine to pull DNA into immature heads. The tail is a large helical structure with a contractile sheath that manages injection of the genome to the host cell. Short tail fibres project from the baseplate. Bent ones are receptor-binding structures.

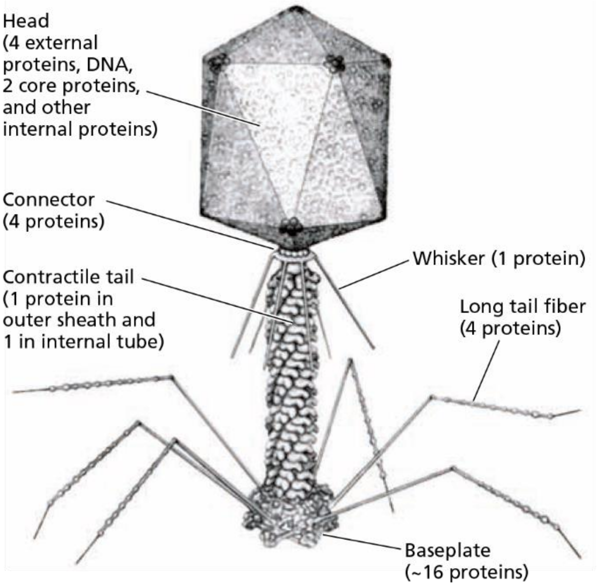


Figure 36. Bacteriophage T4 morphology. A model showing the external components of a common bacteriophage including head and tail. From: Principles of virology 4th Ed. (*250*).

Confirmed oral viruses

The following section contain summaries from the book Oral Microbiology and Immunology 2nd Ed. (*200*)

Several viruses have been detected in the oral cavity, although not for all of them has infection been confirmed. Confirmed viruses in the oral cavity include both DNA and RNA viruses. From enveloped dsDNA: HBV from the *Hepadnaviridae* family, alpha-herpes, *Varicella zoster virus* (VZV), *Herpes simplex virus* 1 and 2 (HSV-1, HSV-2), beta-herpes, CMV, HHV-6, EBV, and gamma-herpes. From non-enveloped dsDNA HPV and MCV. From enveloped ssRNA viruses: *Flaviviruses* such as *Yellow fever virus* (YFV), *Tick-borne encephalitis virus* (TBE), HCV, lentiviruses HIV-1, and HIV-2. From non-enveloped ssRNA, *Foot-and-mouth disease virus* (FMDV).

HBV causes both acute and chronic infections. It is transmitted by contaminated blood or sexual contact. Infection is normally asymptomatic for long periods. Chronic infection increases the chance of liver cancer. It is diagnosed by the presence of HBsAg in serum. It is susceptible to vaccination.

HCV is transmitted from contaminated blood or blood products. It causes chronic infection in more than 70% of all patients, a condition increasing the risk of liver cancer. It is treated with IFN-α and ribavirin.

Viruses from the family *Herpesviridae* are one of the most important groups affecting the oral cavity with HSV-1 causing oral herpes, a recurrent infection and HSV-2 causing genital herpes (STD). Others include VZV causing chicken pox may be detected as well as CMV and EBV which causes severe disease in immunosuppressed patients.

Non-enveloped DNA HPV can cause oral warts or papillomas. HPV 6, 11 have are low risk variants whereas 16 and 18 and have been associated with OSCC are high risk ones. HPV vaccination is reported to prevent the infection of carcinomas. Human p16 overexpression from the CDKN2A gene has been correlated.

HIV-1 and development of AIDS has been detected to favour infection by opportunistic pathogens such as candida and different types of herpesviruses due to a reduced immune response.

The most important cell types of the innate immune system in fighting oral viral infections are macrophages, dendritic cells and NK/NKT cells. B cells can neutralize some of these viruses by binding to their surface. T helper cells recognize viral peptides through the class II presenting pathway. CTLs recognize and kill virus-infected cells through viral peptides presented by the class I pathway. However, viruses may escape the host immune system by inhibiting antigen presentation or as a consequence to mutation, producing decoys and promoting tolerance of the immune system.

The larger picture – Going viral in metagenomics

As stated before, human metagenomics is a burgeoning field that has advanced our understanding of the relationship of the microbiota residing within humans and health. It has also proved capable of identifying its diversity and the way it changes and adapts and there is an ongoing interest in determining the way it may be manipulated. Yet, only the prokaryote fraction has been studied in such fine detail. The same cannot be stated about viruses. The role viruses play in the ecosystem in undeniably important, as described above.

This is understandable, given the technical difficulties viromic studies suppose, even to this date (*263*, *264*). Contrary to prokaryotic metagenomic studies, viral metagenomic projects cannot rely on the amplification of universal molecular markers (e.g. rRNA) as there are no such conserved genes in viruses. Instead, most studies are limited to studying specific groups of viruses or using WGS approaches to explore a wider spectrum, which often does not cope well with taxonomic assignation. Plus, RNA viruses pose a more challenging type of study and databases are still fairly incomplete.

Prior to the onset of the meta-omics era, most viral studies focused on culturable approaches that relied either on growing viruses in host cells *in vitro* or obtaining highly virulent samples that presented a sufficient viral load for sequencing (*265*). Most knowledge in this area was thus built around data from pathogens affecting human or commercially important agricultural harvests and was highly biased towards those species that may be cultured in a laboratory. Classification of some species even responds to morphologic or host-associated similarities rather than to phylogenetics (*260*).

The advent of the meta-omics era saw the total number of known viruses increase exponentially as Rohwer's four ocean expedition and Craig Venter's Sargasso Sea voyage (see Box 3) nearly doubled the existing sequences in the databases in a very short time. It became clear that thousands of species (most of them phages) were still undiscovered or missing characterization (*266*). Viruses no longer needed to be cultured *in vitro* and even some environmental ones with undetectable viral loads were identified thanks to metagenomics whenever sequencing was deep enough (*31*). Yet, only DNA viromes had been analysed (*267*).

There are several reasons why a phylogenetic tree similar to the tree of life cannot be constructed for viruses (*268*): Viruses are polyphyletic and do not have any single gene that is shared genes among all species to trace lineages due to their genomic plasticity. Instead, different proteins may have varying evolutionary origins. Also, viruses are not limited to vertical transmission as they are very susceptible to horizontal gene transfer. This forbids ancestry tracing since much of their genetic contents is obtained from multiple foreign sources, including two-way genetic interchanges with their cellular hosts and they may change the species they infect. Furthermore, similar viruses affecting phylogenetically distant hosts are not necessarily ancient because mechanisms may be adopted horizontally. Furthermore, viral metabolic genes originate from cells, as well as translation genes. Regardless, in 2002, Rohwer and Edwards presented a phage taxonomy tree by using 105 completely sequenced phage genomes available at the time in an effort to create a basis for subsequent studies (*269*). The tree was generated by calculating the aggregative protein distance and penalizing for missing ones (an updated version is found in Figure 37). These authors noted the tree was not intended to reflect evolutionary history but to provide a reference for virus classification dependent on protein profiles. From their analysis, an important observation was that their tree was mostly congruent with the ICTV families and solved some anomalies in current classification system resulting in the reclassification of some families. Phages can be grouped into major clusters Cystophage-Fesselloviridi, Inophage, Microphage, Myophage, Podophage and Siphophage Other proposals for topology reconstructions include employing genomic analyses that focus on structural gene modules by Proux and collaborators (*270*), initial ICTV-based classification as base, and clusters subgroups individually by the Pittsburgh Bacteriophage Institute (*271*), protein folds conservation by Bamford (*272*), protein clusters for identifying open reading frames (ORFs) with no known identity by Yooseph and collaborators (*273*), the gene arrangement in the genome by Li and collaborators (*274*), and prokaryotic Virus Orthologous Groups (pVOGs, previously known as POGs) based on function conservation by Grazziotin and collaborators (*275*). Most of the major groups are relatively conserved among the methods and similar to the ICTV taxonomy (*265*).

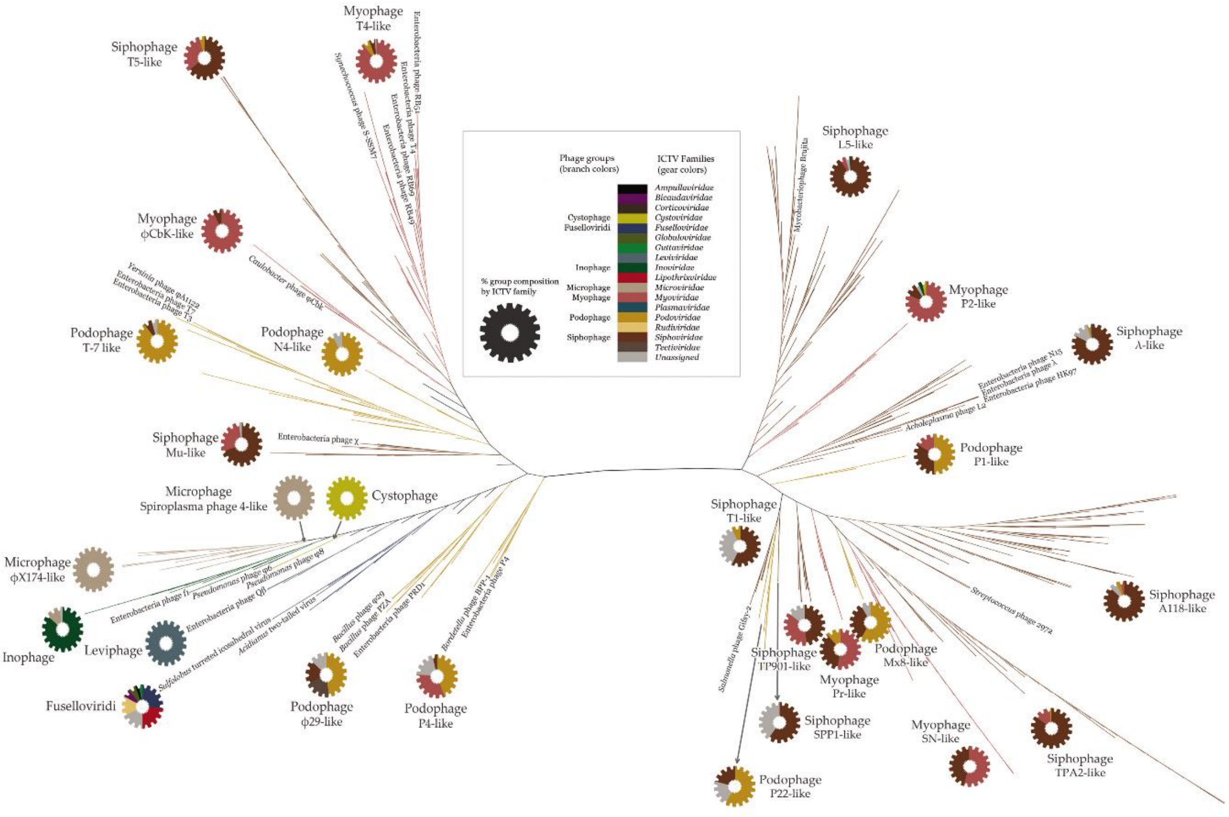


Figure 37. Updated phage protein tree by Rhower and collaborators using protein similarities. A total 1,220 phages infecting Bacteria and Archaea were used for protein prediction and further clustering. The resulting 91,405 groups were aligned and distances were aggregated in distance matrix imposing a penalty for proteins that were not shared. Adapted from: Life in our phage world, 1st Ed (*265*).

Metagenomics has also enabled the study of the relations of prokaryotic and eukaryotic viruses. In prokaryotes the great majority of viruses are dsDNA viruses (and few ssDNA viruses) with genomes with around 10 to 100 kb (*276*). Also, no retroviruses are known. RNA viruses are more diverse in Eukaryotes (mainly positive strand), as well as reverse-transcribing elements and retroviruses that may integrate into the host genome (*277*).

This has led to evolutionary studies of the origins of the major classes using phylogenomics, a very rich topic that escapes the scope of this dissertation. Briefly, current hypothesis points out eukaryotic viruses may have emerged from prokaryotes (*276*). This idea arises from analysing a putative ancestral genome of positive-strand RNA viruses of eukaryotes, which might have assembled from genes derived from prokaryotic retroelements. In particular, dsRNA viruses may have derived from dsRNA bacteriophages or from positive-strand RNA viruses. Their eukaryotic counterparts may have arisen from a fusion of genes from prokaryotic rolling circle-replicating plasmids and positive-stranded RNA viruses. Scientist have also noted that the largest transposons in eukaryotic genomes, Polintons/Maverick, can actually produce virions with infecting capabilities (*278*). This hints on the possibility they may have been an evolutionary intermediate between bacteria and several groups of dsDNA.

The human virome

Phages have been implicated in the control of bacteria in the human body by specializing on species that thrive, reducing their levels back to normal (*267*). This has been extensively detected in the human gut, starting with a work by Breitbart and collaborators in 2002 (*29*). They detected the DNA virome is dominated by viruses from the *Siphoviridae* family and to a lower extend by some from the *Podoviridae* and *Myoviridae* families. The community was estimated to vary greatly in the 1-2 weeks of age. The first human RNA viral metagenome was obtained in 2006 by Zhang and collaborators (*279*). They analysed 18 faecal samples from healthy subjects, reporting the existence of relatively few viruses, mostly comprised by plant infecting RNA viruses Tobacco mosaic viruses (TMV) and Pepper mild mottle virus, (PMMV). They confirmed their presence with RT-PCR in samples from America and Asia, assuming it was highly prevalent in the human population. Precisely, TMV was in fact found in a study of the saliva of smoking subjects (not found in control non-smokers) (*280*). As happens with most viruses from non-culturable studies, its actual role in health and disease is poorly understood, although some studies have correlated its detection in stool samples with fever, abdominal pain and pruritus (*281*), an observation that other authors have criticized, indicating it as just a possibility that must first be confirmed with infection demonstration, the same as for other plant viruses, such as tospoviruses, rhabdoviruses, reoviruses, begomoviruses and nanoviruses.

In a classical study by Reyes and collaborators of the virome of monozygotic twins and their mothers using pyrosequencing the authors compared the composition of their bacterial communities (*63*). Most were *Podoviridae*, dsDNA bacteriophages and many were thought to be prophages. Eukaryotic viruses such as *Herpesviridae* and some RNA viruses (*Reoviridae*) were found as well. Twins and their mothers presented a significantly similar virome. They also reported a greater interpersonal variability in the viral communities (as well as functionally) when compared to their bacterial counterpart. However, over 90%

of the most common viruses were retained over time. They suggested this was due to the bacterial population being stable.

Although much was known by the early 2010s about the microbial community in niches such as the mouth, viromics were just starting. An example of this was a study by Willner and collaborators in which they studied DNA viral communities in the mouth using oral swabs (*282*). They found vast amounts of phages, including a T3-like phage that was almost completely sequenced. Also found were *Propionibacterium acnes phage* PA6 and *Streptococcus mitis phage* SM1. The later was significant as *S. mitis* plays a role in oral pathogenesis. Fragments of EBV homologous sequences were also recovered. A trait that matched results in other studies was the large proportion of unknown results (those with no hits in databases) and bacterial genes in the samples which presumably are detected due because of technical caveats.

Clinical researchers have also done viral metagenomics using alternative approaches, such as in a study published the group by David Pride in 2012 (*283*) using IonTorrent sequencing. These authors analysed saliva from five healthy subjects over a 2-3 month period. Similar to Willner's results, the data revealed viral communities in saliva were dominated by bacteriophages. Comparison with bacteria showed concordance among certain predicted pairing of hosts and viruses but not all, suggesting some may in fact be related previously unreported phage activity towards other bacteria. Most viral sequences were homologous across all timepoints within each subject and there was larger variability between subjects than among several samples from the same subject meaning profiles were probably due to inherent properties of each patient. The authors also reported some virulence factor in viral involved in immune evasion through breakdown of complement or IgA adhesion. Moreover, subjects living in the same household had more similar profiles. In a subsequent study (*284*), the same research group analysed the oral viral community in association to living environment (different households). Not all groups living together presented similar communities. An additional strategy was using clustered regularly interspaced short palindromic repeat (CRISPRs), viral signatures in bacterial mechanisms used to recognize invading viruses. By analysing these for five bacterial species they managed to determine members of a single household had been exposed to similar viruses. On a longer longitudinal study (days 1, 2, 4, 7, 14, 30 and 60) with saliva from eight patients, the group found that some of the viruses were shared within the individuals at different timepoints, as well as between unrelated individuals (*285*). More recently, Pride's group analysed viral communities in saliva, subgingival and supragingival biofilms. They detected differences in composition were significantly associated with the status of the plaque. They detected an increased amount of myoviruses in the subgingival biofilm that they attributed to lytic phages (*286*).

Most novel viral knowledge is still established by culture-free amplicon typing (sequencing and amplifying targeted viruses) in a small scale. As some examples, a new genus of circular ssRNA, *Cyclovirus* was detected in the human gut (*287*), same as particles from the *Polyomaviridae* family, that were thought to be transmitted via a faecal-oral route (*288*) and ssRNA *Noroviruses* from the gut of healthy symptomatic and asymptomatic children (*289*). All of them were directly found by PCR (retrotranscriptase-PCR, RT-PCR, in the case of RNA viruses).

As stated before, it has been discovered that TTV and other *Anelloviridae* are present in human blood in healthy patients (*246*). Further studies have confirm them to be widespread commensals of most primates and found in high prevalence in the human population (>80%), located mostly in the bone marrow, lung, spleen and liver (*290*). Also, as an example of human virome in the respiratory system, Wylie and collaborators carried out a study of children with unexplained fever using nasal swabs. They found *Dependovirus* and *Bocavirus*, from family *Parvoviridae* in the healthy controls but at lower quantities in affected children, though a beneficial role could not be confirmed (*291*). They also detected paramyxoviruses and adenoviruses in healthy children, *Influenza A virus*, *Coronavirus*, etc., some of which are associated to disease. As can be deduced from these and many others studies, the human body has been thoroughly scanned for viruses (Figure 38), with most studies focusing small sets using PCR amplifications and some important, yet mostly disease-oriented viral metagenomic studies.

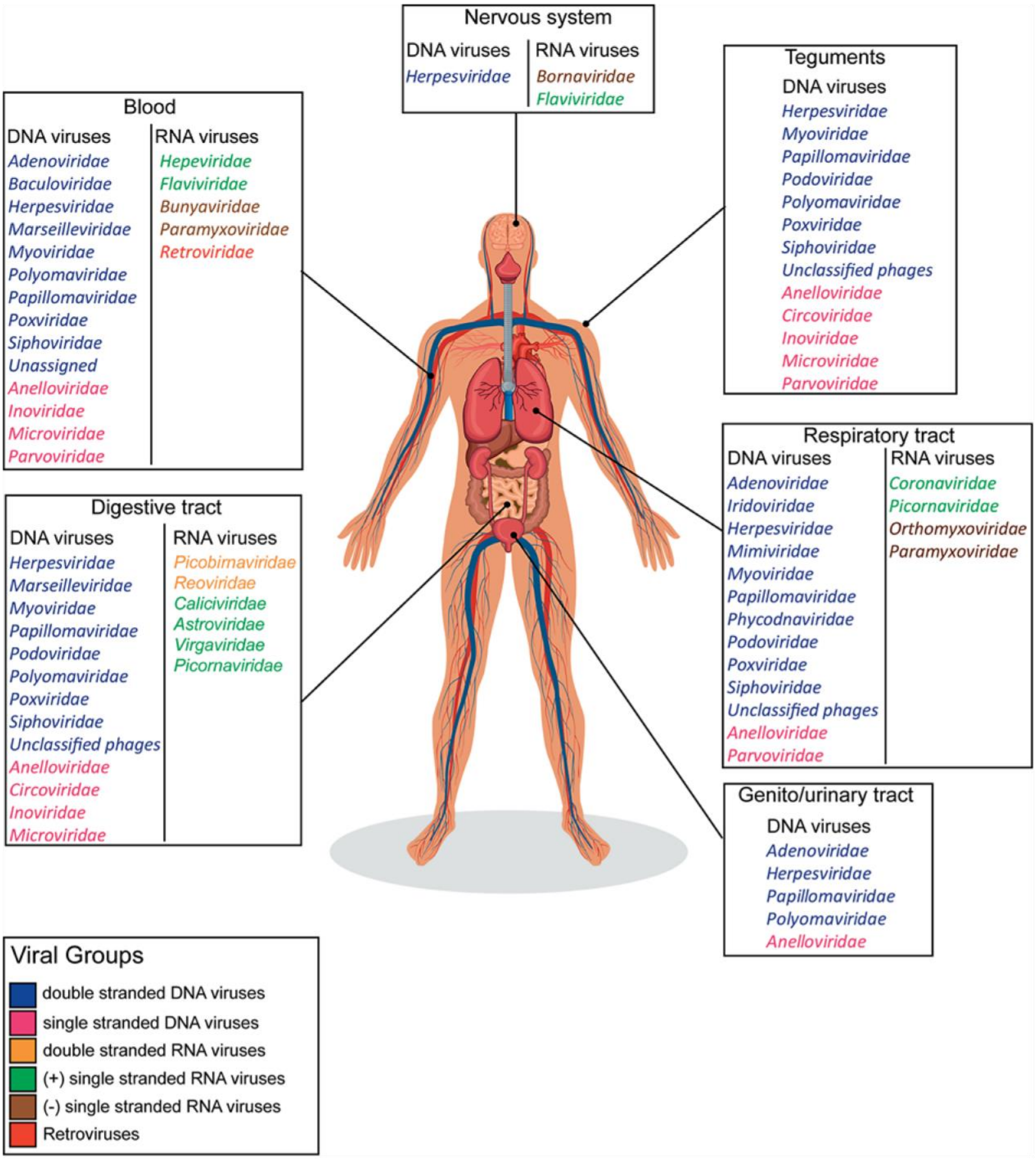


Figure 38. The human virome in non-pathogenic conditions. Distribution is presented for the most relevant viral families found in the major human systems. From: Popgeorgiev *et al.*, 2013 (*292*).

The viral component of the HMP was analysed by Wylie and collaborators focusing in the search of dsDNA Eukaryotic viruses (*285*). They used 706 samples from 102 subjects with each sampled at the nose, skin, supragingival plaque and stool (52 individuals were sampled at 2-3 timepoints 30 to 359 days apart). Although the HMP data was focused on prokaryotes and the method was not exhaustive, this secondary study resulted in the largest viromic analysis of the human body, achieving the detection of an average 5.55 viral genera in each individual with at least one detected in 92% of all individuals. Most prevalent were herpesviruses, papillomaviruses, polyomaviruses, adenoviruses, anelloviruses, parvoviruses and circoviruses. Each individual presented a distinct profile and some were stable over time and profiles were different between body sites. For instance, the most commonly detected genus in the mouth were *Roseolovirus, Lymphocryptovirus,* and *Betapapillomavirus* and was the only niche where circoviruses were detected, a result that was at least partially consistent with previous analyses. Curiously, fewer viruses could be recovered from the stool samples. The skin presented the most genera, followed by the nose. Also, the skin contained more papillomaviruses than any other niche.

In 2016, Paez-Espino and collaborators published a large-scale reanalysis of over five Tb of metagenomic sequence data from 3,042 geographically diverse samples from previously available studies (*293*). They managed to predict 125,000 partial viral genomes by training recognition patterns into an algorithm for searching the whole database, increasing by 16-fold the number of known viral genes. Also, using CRISPR spacers and tRNA matches they manage to map many of the bacteria to their potential predators. They also reported that their analysis suggested more than 30% of intestinal and 50% of oral viral sequences were shared by at least 10% of the sampled subjects and estimated there were an average of 3.4% and 7.4% of viral sequences in all the oral and stool samples, respectively, from the HMP (a major data source for the study), more than previously estimated.

Some details on techniques for studying the microbiota

The expansion of automated molecular biology technologies has promoted a gradual shift from the paradigm of single-species culturable genomics to a richer, yet defiantly more complex, metagenomic approach. Even so, evolution of technology has always been reciprocal and has successfully adapted the requirements of the scientific community into its projections for the future, effectively shaping newer technologies to better respond to contemporary biological questions. Since metagenomics depends deeply on the generation of massive data, PCR amplification and high-throughput sequencing platforms have consolidated as the stepping stones of hardware technology for this burgeoning science. Sensibly, the disproportionate growth of datasets has been toppled by current computational limitations for the analysis of the immense loads of generated data, thus requiring robust and optimized bioinformatic approaches to tackle the problem. For the rest of the manuscript only techniques relevant to the presented work will be commented.

The study of the microbiota has change considerably since it first started, back in the early 2000s. Different molecular and bioinformatical approaches have been introduced in the last decade and a half to study the microbiota, evolving throughout more than two very busy decades. Most notably, sequencing platforms have been adapted and informatic modules have been tailored for the study of the different nucleic acids in metagenomics and metatranscriptomics (RNA-seq) studies exploring genomic DNA and

RNA transcripts(*64*), respectively. Nowadays, the scope of the studies has expanded towards the analyses of higher level multi-omics such as metaproteomics and metabolomics (*64*).

Sequencers from the second generation of automated platforms, such as Illumina MiSeq and HiSeq operate by elongating multiple-copy DNA templates in very large amounts (*47*) and are susceptible to amplification biases. The third generation platforms use a single-molecule sequencing approach producing long reads but are instead affected by low quality base calling (*54*, *56*).

Metagenomics traditionally explores all the genomic DNA (gDNA) (*13*) from all species found in an environmental sample (whole genome shotgun sequencing, WGS) whereas for metatranscriptomics whole collections of messenger RNA molecules (mRNA) are translated into cDNA for sequencing(*294*). The former gives an insight into the functional potential of the microbiota whilst the latter is employed to assay genes being translated, and thus the active fraction of the microbiota.

The resulting sequences (reads) from each sample are normally bioinformatically processed to be identified by aligning them or using their genomic features to associate them to items in a database that hopefully contains potential homologs (*295*). Genetic or functional annotations can then be drawn from such related sequences to rebuild the potential profiles for a sample and be statistically compared against others.

WGS metagenomics, however, provides only limited information about the taxonomic content of a sample (*296*). This is because this approach is a randomized process based on the intrinsic abundance of the genomes in a sequencing library. Different genomes appear in varying proportions and their contents are sequenced unevenly with the most abundant and readily accessible molecules or copies having a higher probability of getting sequenced (*32*). Furthermore, there is much redundancy in metagenomics libraries (especially at more conserved regions of the genomes).

This is why, for taxonomic exploration of prokaryotes, most studies rely on the existence of universal molecular markers with predictable variability at different taxonomic levels (*82*), the most common being the highly conserved rRNA gene (also referred to as rDNA). The sequencing of the small subunit of the rRNA gene provides a fairly accurate way of identifying archaea and bacteria (16S) or fungi (18S). The internal transcribed spacer (ITS), the region separating the small and large subunits of the ribosome, is also commonly used for taxonomic identification of fungi. In recent years, the number of 16S profiling studies for exploring the prokaryotic composition and abundance has seen an unprecedented growth as sequencing costs have become more accessible (Table 7). This decline responds to the release of new platforms with higher sequencing throughput and longer sequence lengths, optimizing the costs of price per read. Bacteria are by far the most studied group of the microbiota and have the greatest number of available tools for their analysis.

Table 7. Comparison between Illumina and 454 platforms. From: All-Seq Inc. (*50*)

	Illumina							**Roche 454**	
	HiSeq X Ten*	**Hi Seq 2500**			**NextSeq 500**				
		HT v4	*HT v3*	*Rapid*	*High*	*Mid*	*MiSeq*	*GS FLX+*	*GS Jr.*
Total output	1.8 Tb	1 Tb	600 Gb	180 Gb	129 Gb	39 Gb	15 Gb	700 Mb	35 Mb
Run time	3 days	6 days	11 days	40 hrs	29 hrs	26 hrs	~65 hrs	23 hrs	10 hrs
Output/day	600 Gb	167 Gb	55 Gb	~110 Gb	~100 Gb	~36 Gb	~5.5 Gb	700 Mb	35 Mb
Read length	2 X 150 b	2 X 12 b5	2 X 100 b	2 X 150 b	2 X 150 b	2 X 150 b	2 X 300 b	up to 1 Kb	~700 b
# of single reads	6 B	4 B	3 B	600 M	400 M	130 M	25M	1 M	0.1 M
Instrument price	$1 M*	$740 K	$740 K	$740 K	$250 K	$250 K	$125 K	~$500 K	$125 K
Run price	~$12 K	~$29 K	~$26 K	~$8 K	$4 K	?	~$1.4 K	~$6 K	~$1 K
$/Gb	$7	$29	$43	$44	$33	?	$93	$8.5 K	28.6 K

* HiSeq X Ten* consists of an array of ten HiSeq sequencers

The 16S rRNA gene presents six total hypervariable regions over a span of ~1,500 nucleotides (it varies depending on the species), each with differential taxonomic resolution capability (*37*). Yet, limitations in current technologies allow for the sequencing of only 2-3 consequent regions, rendering resolution dependent on the specific regions of choice (Figure 39). The 16S profile is obtained by sequencing amplicons produced by polymerase chain reaction (PCR) amplification using primers flanking the desired hypervariable regions. Thus, even though 16S is the most popular technique, 16S profiling is far from perfect as intrinsic PCR biases apply. A very extended rule of thumb indicates two 16S sequences that come from the same species should share a sequence identity of over ~97%(*297*), whereas two sequences from the same genus share ~95%. In practice, this markedly depends on the resolution of the sequence that is amplified. Some more accurate estimates for the whole sequence are found in Table 8.

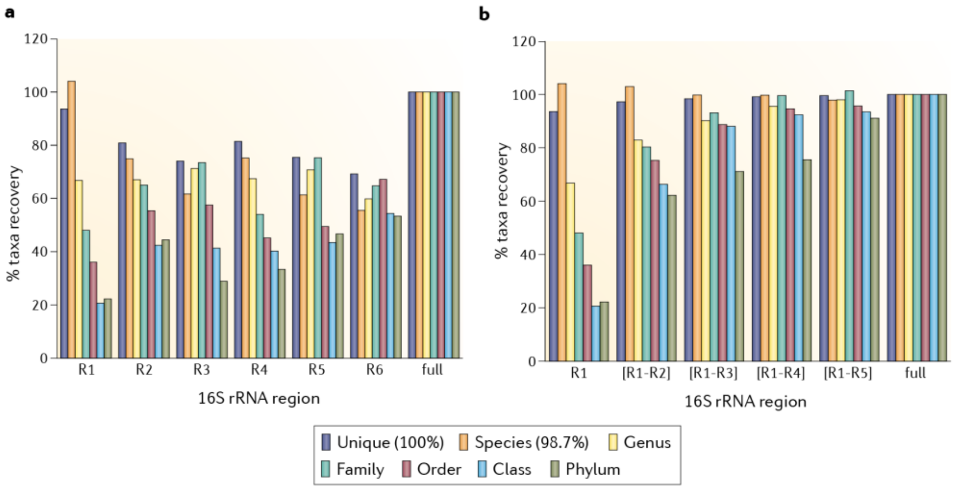


Figure 39. Predicting taxa richness using partial 16S ribosomal RNA sequences. A) Six fragments of 250 nt that differentially capture diversity in the samples at different taxonomic levels. The R1 segment includes hypervariable regions V1 and V2, R2 includes V3, R3 includes V4, R4 includes V5 and V6, R5 includes V7 and V8, and R6 includes V9. B) Longer fragments of 250, 500, 750, 1050 1300 and full-length comparison. From: Yarza *et.al.*, 2014 (*37*).

Table 8. Taxonomic thresholds of the 16S rRNA gene in bacteria and archaea. From: Yarza *et.al.*, 2014 (*37*).

	Number of taxa	**Median sequence identity %**	**Minimum sequence identity%**	**Threshold sequence identity %**
Genus	568	96.4 (96.2, 96.55)	94.8 (94.55, 95.05)	94.50
Family	201	92.25 (91.65, 92.9)	87.65 (86.8, 88.4)	86.50
Order	85	89.2 (88.25, 90.1)	83.55 (82.25, 84.8)	82.00
Class	39	86.35 (84.7, 87.95)	80.38 (78.55, 82.5)	78.50
Phylum	23	83.68 (81.6, 85.93)	77.43 (74.95, 79.9)	75.00

Confidence intervals are presented in parenthesis.

Most importantly, analyses depend on the initial nucleic acids as protocols vary greatly for RNA methods. Also, large amounts of sequences are needed because viruses do not have universally conserved sequences such as the rDNA gene for evaluating taxonomy. Thus, only through WGS can the virome be explored. Furthermore, there is need for amplification since commonly, unless the population in the samples is going through an exponential phase, may not enough DNA may be recovered. Finally, viral sequence databases are incomplete, limiting the identification range.

Understandably, the viral fraction is a mostly neglected part of metagenomics, mainly because of several technical and sequence-related complications that impairs its exploration: Viral do not reproduce predictably as do bacteria and thus the number of total viral particles obtained during sampling may be very small (*250*). Viruses were initially obtained by filtering large amounts of liquefied samples to separate the viral fraction (*256*). This secured a large enough number of viral particles for DNA to be sequenced. Current techniques allow for the study of much smaller quantities via different enrichment protocols (*264*). The genome of multiple viruses is coded in RNA, rendering retrotranscription unavoidable for sequencing, which is similar to studying a metatranscriptome without the added benefit of having a poly(A) tail of messenger RNA. There is no universally conserved marker such as the rRNA molecule found across all viruses, which makes WGS the only approach for sequencing (*267*). The success therefore depends highly on the depth of sequencing and inherent underlying diversity of the virome.

This is particularly problematic since assembling viral metagenomic sequences without reference (*de novo* assembly) is further complicated by the high amount of HGT that occurs within this group, which may result in the formation of chimeric contigs (*298*). They may import foreign DNA into their own genomes and some of them are found within bacterial genomes in the form of prophages (*256*), further complicating taxonomic classification.

Classification in viruses is utterly challenging as some older classification was done with morphological features or host-association (*260*). Also, taxonomic levels in viruses are not standardized (*259*). Most of them have a family, genus and species and some have an order or subfamily, although it is common for them to be missing two or more levels (*298*). Baltimore's classification uses the source nucleic acids instead but is not officially recognized as a category. To make matters worse, most databases are incomplete and mistakes are abundant (*298*). Bad sequences annotation is extended in all official databases and they are usually carried when a new study does homology searches using such registries or classified databases as their source. Also, viral databases are highly biased towards plant and human pathogens. Important efforts have been made to filter sequences and species-specific databases are often curated to address these issues but still, their main problem is that most species are probably still missing. In fact, most viromic projects still have a large proportion of unidentified sequences owing to the scarce knowledge of the real variability of viruses in the environment (*267*).

HYPOTHESIS AND OBJECTIVES OF THE THESIS

Proliferative verrucous leukoplakia (PVL) is considered a high-risk variant of oral leukoplakia (OL) characterized by the initial formation of asymptomatic hyperkeratotic patches that may evolve into exophytic wart-like protrusions and, eventually, oral squamous cell carcinoma (OSCC). PVL lesions have been described as slow-progressing but their resilience and high malignancy potential underline the importance of its diagnosis. Detection of PVL in its early stages is complicated as the lesions resemblance simpler forms of leukoplakia. Histopathological examination and the observations of multifocality and recurrence are required for diagnosis confirmation. No aetiological agent has been confirmed for PVL but the appearance pattern and type of lesions suggests a viral agent could be involved.

As an environment, the oral cavity is inhabited by hundreds of thousands of microbes, including numerous species of bacteria and viruses. Together, they form a complex community: the oral microbiota that actively interacts with its host. The study of the bacterial fraction (microbiome) and the viral fraction (virome) has been enabled by technological and scientific advances in the field of metagenomics. These and other meta-omic techniques for culture-independent large-scale analyses survey the whole spectrum of environmental organisms and particles to elucidate their role and composition, as well as the effect they may have in health and disease.

The main objective of the present dissertation is the identification of a putative aetiological agent of proliferative verrucous leukoplakia via the exploration of its associated microbiome and virome in a retrospective exploratory study.

Hypothesis: The aetiology of PVL resides in an infectious agent, possibly a virus, that is part of the lesion-associated microbiota.

Objective 1: Discarding oncogenic viruses as potential aetiological agents.

In most cases, the high malignancy of PVL leads to the development of OSCC as the lesion evolves. Previous studies from the scientific community have tested the existence of oncogenic DNA viruses: Human papillomaviruses (HPV) and Epstein-Bar viruses (EVB), obtaining mixed results, but none were confirmed as aetiological agents. For this objective, we carried out a retrospective screening of known human oncogenic viruses using Polymerase Chain Reaction (PCR) amplification to evaluate its presence in samples from patients with OL, PVL, OSCC and healthy controls.

Objective 2: Constructing adequate bioinformatic tools for studying the metagenomics datasets.

The analysis of the bacterial fraction of the microbiota is well standardized as most metagenomic studies focus on it. Virome analyses, on the other hand, face several limitations. Most importantly, there are no markers conserved in all viruses, retrotranscription is required for sequencing RNA viruses, and most sequence databases containing viral references are small or lack adequate taxonomic information.

For this objective, bioinformatic scripts were programmed to tackle the various technical challenges of viral metagenomic analysis and suitable databases were accordingly constructed.

Objective 3: Studying the bacterial and viral composition in search of an aetiological agent.

Taxonomic profiles provide an insight into the composition and abundance of environmental samples, which can be used to assess significant variation among samples and between different groups in a study, identifying major players or conditions that may be correlated to certain species configuration. Bacterial diversity can be analysed by PCR amplification of the 16S rDNA gene, followed by high-throughput sequencing of the amplicons (16S profiling), and assignment using 16S references of representative taxonomic groups. Viral diversity can be estimated from whole genome shotgun sequencing (prior retrotranscription of the RNA fraction) and subsequent identification using genomics features or genomic references. For this objective, we analysed and compared viral and bacterial taxonomic and functional profiles from samples from OL, PVL, OSCC and controls to identify potential markers for each group and its most important correlations. An extension of the WGS viral analysis was carried out to survey metagenomic datasets for other groups of organisms.

MATERIALS AND METHODS

A) EXPERIMENTAL PROCEDURES

Oncogenic DNA viruses screening in three groups of lesions and controls

In order to confirm or discard currently known oncogenic DNA viruses as aetiological agents prior to a larger metagenomic analysis, a retrospective study was designed as a blind experiment to evaluate the presence of their PCR screening. The results from this screening have been published in the journal *Clinical Otolaryngology* (*210*).

Participants and sample collection

Samples from patients attending the Stomatology Unit in Hospital General Universitari in Valencia are regularly collected as part of the clinical procedures and examination after signing informed consent and are stored at -80ºC. Personal identification for clinical studies is anonymized. For this study, samples from 40 subjects with an average of 58.20±21.73 years of age were selected from, 72.50% from female patients. The study was approved by the local research ethics committee.

Biopsies from four groups of patients were collected, each comprised of ten subjects. The diagnosis of patients participating in the screening was determined by clinical assays followed by the extraction of an incisional biopsy carried out using local anaesthesia. Biopsies were split in two parts by the dental surgeon. Half of each sample was destined to conventional histopathological studies whilst the other was stored in RNAlater Solution (Ambion) at -80º C for later analysis. The first group included samples from oral leukoplakia (OL) lesions (age mean = 65.50±9.86, 90% female). The second was composed of biopsies from patients that matched the diagnostic criteria for proliferative verrucous leukoplakia (PVL) as defined by Cerero-Lapiedra and collaborators (*211*)(age mean = 70.30±8.12, 100% female). The third one was comprised from patients with oral squamous cell carcinoma (OSCC) but not presenting PVL characteristics (age mean = 72.10±14.11, 30% female). Finally, a control group consisted of biopsies from patients with no related precancerous nor PVL lesions that were collected during the extraction of third molars (age mean = 24.90±3.98, 70% female). The pathological status of the OSCC, PVL and OL was confirmed by histopathological studies of the biopsies.

The fractioned biopsies were once again cut in half using sterile scalpels while still frozen. One part was immediately transferred to 1.5ml microcentrifuge tubes (Eppendorf) for extraction while the rest was stored at -80º C. The fragments were transferred to new tubes after being rinsed with 300 µl of Hank's Balanced Salt Solution (HBSS) buffer (GIBCO) at 4ºC to wash off the RNAlater solution as it proved to be an issue during trial extractions for procedure optimization. A volume of 600 µl HBSS was added to each tube.

DNA Extraction

In order to break the tissue in hyperkeratotic patches for better homogenization, sample structure was mechanically disrupted using a Tissue Lysser II (Qiagen) (30 Hz, 5 to 10 minutes depending on its hardness) with sterile 3 mm tungsten carbide beads (Qiagen). During the process, sample temperature was kept as low to as possible by previously cooling down the Tissue Lysser II adaptors at -20ºC as friction elevates temperature.

A phenol/chloroform DNA extraction was carried out using a conventional protocol. Briefly, for each sample, a volume of 30 µl phosphate-buffered saline (PBS) buffer 10% was added. The solution was incubated for 15 min at room temperature and 170 µl of Tris-$MgCl_2$-NaCl buffer (TMN) buffer were added. After a volume of chloroform was added, the tubes were inverted and centrifuged at 3000x g for 5 min at room temperature. The aqueous phase was carefully collected and transferred to a new tube; the remaining volume was discarded. The same procedure was repeated with an equal volume of phenol, with a 1:1 mixture of phenol-chloroform, and finally with a 24:1 chloroform isoamyl alcohol mixture (AppliChem).

DNA was precipitated by adding 40 µl of NaOAc 3M and 1ml EtOH 100% at -20ºC. The tubes were incubated in ice for 15 min and centrifuged at 12,100 x g for 3 min. The supernatant was discarded the pellet was cleansed twice by adding 1 ml EtOH 70%, vortexing and centrifuging at 12,100 x g for 5 min. The pellet was air dried at room temperature for 10 min. DNA was resuspended in 50 µl TE buffer (1 M NaCl, 1 M Tris-HCl, pH 8.0, 0.5 ethylenediaminetetraacetic acid (EDTA).

In order to confirm that the extraction had been successful, co-extracted bacterial DNA within the samples was used as positive control amplified using a PCR amplification targeted at the V1-V2 hypervariable regions of the 16S rDNA gene with forward E8F (5′-AGAGTTTGATCMTGGCTCAG-3′) and reverse B530R (5′-CCGCGGCKGCTGGCAC-3′) primers. The PCR were carried out using a Biomix^{TM} PCR kit (Bioline) in a Mastercycler Pro thermocycler (Eppendorf). PCR conditions were as follows: denaturation cycle at 95ºC for 2 min; 30 cycles of denaturation at 95ºC for 30 s, annealing at 52ºC for 1 min, and extension at 72ºC for 1.5 min; final elongation step at 72ºC for 10 min. All pre- and post-PCR experiments were carried out in separate areas to prevent contamination. Blank controls were systematically included to monitor contamination events.

The presence of control amplicons was confirmed by running 1.4 % agarose gel electrophoreses (550-650 bp band). Briefly, samples were stained with GelRed^{TM} nucleic acid gel stain (Biotium) and detected under a UV light. For those where no bands were detected, the extracted DNA was enriched with Illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare Life Sciences) following the manufacturer’s instructions. DNA concentration was measured with Quant-iT™ PicoGreen ® dsDNA Reagent (Invitrogen) and 60 µg/ml DNA were loaded in a new round of PCR to amplify 16S rDNA. This second iteration was carried out using the same PCR conditions except for cycles. Only samples that yielded amplified products of bacterial DNA were considered for subsequent analyses.

PCR amplification

The main objective of the experiment was the PCR amplification for detection of DNA oncoviruses *Human Papillomavirus* (HPV), *Human Polyomavirus* (HPyV) including *BK virus* (BKV) and *Merkel Cell*

Polyomavirus (MCP), and *Human Herpesvirus* (HHV), including *Kaposi's sarcoma-associated virus* (HHV-8) and Epstein-Bar Virus (EBV). Sets of primers for each group were obtained from the literature and were tested bioinformatically, simulating amplifications *in-silico* with PrimerProspector (*299*) and Reference sequences from NCBI's RefSeq.

For HPV, two sets of degenerate primers were used: Pairs FAP59: 5'-TAACWGTIGGICAYCCWTATT-3' and FAP64: 5'-CCWATATCWVHCATITCICCATC-3' (*300*), and pair CP4: 5'- ATGGTACARTGGGCATWTGA-3' and CP5: 5'-GAGGYTGCAACCAAAAMTGRCT-3' (*301*), both designed to recover over 60 HPV types including high-risk types (e.g. HPV-16 and HPV-18). The PCR conditions for both sets were as follows: denaturation cycle at 94ºC for 5 min; 35 cycles of denaturation at 94ºC for 15 s, annealing at 45ºC for 1 min, and extension at 72ºC for 20 s; final elongation step at 72ºC for 10 min. DNA from *Eidolon helvum papillomavirus* was included as a positive control for both sets of primers. The amplification of a fragment of ~500-600 bp was expected.

Polyomaviruses were surveyed with sets of degenerate primers designed for amplification of HPyV related to MCP and BKV using nested PCR (*302*). For the first PCR, forward primer VP1-1f: 5'-CCAGACCCAACTARRAATGARAA-3' and reverse VP1-1r: 5'-AACAAGAGAACACAAAT(N/I)TTTCC(N/I)CC-3' were used. The conditions were: denaturation cycle at 95ºC for 12 min; 45 cycles of denaturation at 95ºC for 30 s, annealing at 46ºC for 30 s, and extension at 72ºC for 2 min; final elongation step at 72ºC for 15 min. For the second reaction, forward VP1-2f: 5'-ATGAAAATGGGGTTGGCCC(N/I)CT(N/I)TGYAARG-3' and reverse VP1-2r: 5'-CCCTCATAAACCCGAACYTCYTC(H/I)ACYTG-3' primers were used (~250-300 bp amplicon). The conditions were: denaturation cycle at 95ºC for 12 min; 45 cycles of denaturation at 95ºC for 30 s, annealing at 50ºC for 30 s, and extension at 72ºC for 2 min; final elongation step at 72ºC for 15 min. An additional primer set specifically designed for MCV in carcinoma samples was selected (*303*). The forward primer was MCVPS1f: TCAGCGTCCCAGGCTTCAGA and the reverse was MCVPS1r: 5'-TGGTGGTCTCCTCTCTGCTACTG-3'. The conditions were: denaturation cycle at 95ºC for 5 min; 35 cycles of denaturation at 95ºC for 30 s, annealing at 55ºC for 30 s, and extension at 68ºC for 90 min; final elongation step at 72ºC for 7 min. No positive control was available for us at the time the HPyV screening was carried out.

The last group of viruses that was targeted was the HHV. The screening was carried out using sets of degenerate primers specifically designed for universal herpesvirus amplification (22 species, including HHV-8 and EBV) with nested PCRs (*304*). For the first round of PCR, two forward primers were used: DFAf: 5'-GAYTTYGCNAGYYTNTAYCC-3' and ILKf: 5'-TCCTGGACAAGCAGCARNYSGCNMTNAA-3' and a reverse KG1r: 5'-GTCTTGCTCACCAGNTCNACNCCYTT-3'. The second PCR, the inner amplification, was carried out using forward primer TGVf: 5'-TGTAACTCGGTGTAYGGNTTYACNGGNGT-3'and reverse IYGr: 5'-CACAGAGTCCFTRTCNCCRTADAT-3'. PCR conditions for all reactions were: denaturation cycle at 94ºC for 5 min; 45 cycles of denaturation at 94ºC for 30 s, annealing at 46ºC for 1 min, and extension at 72ºC for 60 s; final elongation step at 72ºC for 7 min. The positive control contained DNA from EBV and HHV-8, expected to produce a ~600 bp fragment after amplification.

Amplified PCR products were run in a gel electrophoresis. Any bands detected in the agarose gel were considered for Sanger sequencing. A representative sample for each band was selected for the sequencing reactions (generally the ones yielding the most intense and well defined bands in the agarose gel). Whenever possible, two samples sharing similar size bands were selected. A volume of 8 µl was cleansed using a NucleoFast® 96 well PCR clean-up plate (Macherey –Nagel) following the manufacturer's instructions to remove undesired PCR products. Samples in which two bands were detected were loaded in an agarose gel (0.8%) and the resulting bands were excised from the gel. These were subsequently cleansed using a High Pure PCR Product Purification Kit (Roche) according to the manufacturer's instructions.

Sequencing and informatic processing

The sequencing reactions were carried out with a BigDye Terminator v3.1 Cycle Sequencing Kit (Life Technologies) by following the manufacturer's instructions. Sanger sequencing was carried out in a 3130xl Genetic Analyzer (Applied Biosystems) from the Universitat de València SCSIE Genomics Facility at the University of Valencia (Spain). The obtained sequences (reads) were post-processed using the Staden Package for quality trimming (*305*). The resulting sequences were submitted to the DDBJ Nucleotide Sequence database (AB897841-AB897864). The BLAST+ v2.2.27 suite from NCBI (*306*) was used for homology searches . Depending on the primers that were used, each group of reads was aligned to its corresponding viral family subset (*Papillomaviridae, Herpesviridae* or *Polyomaviridae*) from NCBI's non-redundant database (nr database) (downloaded on Sept 11, 2013), using local megablast setting the evalue threshold to 1e-5 to identify highly similar sequences. Sequences bearing no significant matches were then aligned to the same subsets with tblastx with evalue >1e-3 in search of more distant sequences with conserved protein structure. This algorithm was selected because it considers all six reading frames in the query sequence for protein translation and compares each expected amino acid sequence against all possible translations of sequences in the database. Sequences that had not been assigned taxonomy placement were then compared against the whole nr database using blastn with evalue of 1e-3 in order to look for remote homologous DNA sequences.

<u>Exploration of the microbiome and virome in three groups of lesions and controls</u>

In order to search for an aetiological agent of PVL within the underlying microbiota, a new set of samples was obtained for a retrospective exploratory analysis of the viral (DNA/RNA) and microbial using culture-independent approaches and high-throughput sequencing. The study followed a similar recollection scheme to that of the screening study comparing different type of lesions.

For bacteria, 16S rDNA profiles were obtained for pyrosequencing whereas for viruses and bacteria, whole genomic sequences were prepared for Illumina sequencing.

Participants and sample collection

A new set of samples stored at -80ºC from patients attending the Stomatology Unit in Hospital General Universitari in Valencia was selected. These patients had been informed and the data for clinical studies was anonymized. For this retrospective study, samples from 41 subjects with an average of

74±16.11 years of age were selected from the Hospital General Universitari in Valencia, 58.54% were female. The study was approved by the local research ethics committee.

Incisional biopsies from four groups of patients were collected: PVL, OL, OSCC and controls (controls were eleven). Again, the diagnoses were determined by clinical assays and histopathological studies, which used half of the sample whilst the other was stored in RNAlater Stabilization Solution (Ambion) at -80º C for later analysis. The first group included ten samples from oral leukoplakia (OL) lesions (age mean = 60.30±17.10, 60% female). The second was composed of ten biopsies from patients that matched the diagnostic criteria for proliferative verrucous leukoplakia (PVL) as defined by Cerero-Lapiedra and collaborators (*211*)(age mean = 70.40±11.84, 70% female). The third one were from 16 patients with oral squamous cell carcinoma (OSCC) but not presenting PVL characteristics (age mean = 77.31±14.36, 50% female). Finally, a control group consisted of five biopsies from patients with no related precancerous nor PVL lesions (age mean = 23.40±1.13, 60% female).

Nucleic acid extraction

Biopsies were transferred to 1.5 ml microcentrifuge tubes (Eppendorf) and rinsed with 300 µl of HBSS buffer (GIBCO) at 4ºC to wash off the RNAlater and they were transferred to 2 ml tubes. Most samples were 3-7mm long so samples larger than that were cut to 5 mm using sterile scalpels. A volume of 600 µl HBSS was added to each tube. Tissue structure was mechanically disrupted using a Tissue Lysser II (Qiagen) (30 Hz, 5 to 10 minutes depending on its hardness) with sterile 3 mm tungsten carbide beads (Qiagen) at 4ºC. The samples were centrifuged at 4,000g for 5 min and the supernatant was transferred to new 2ml tubes. The pellet was stored at -80ºC for subsequent bacteria extraction.

For viral DNA/RNA extraction. The supernatants were serially filtered through cellulose syringe filters with pore sizes 5.0 µm, 0.8 µm, 0.45 µm and 0.20 µm to eliminate bacteria (Sartorius Stedim Biotech). As the nucleic acid contents of viruses is expected to be protected within capsids, the rest of the DNA (from bacterial and human origin) is treated using a cocktail of DNases and RNases consisting of 14 U of TURBO^{TM} DNase (2 U/ µl) (Ambion), 20 U of Benzonase® Nuclease (Novagen) and 20 U of RNase A (Invitrogen) in DNase Buffer (Ambion), 120 min at 37ºC. Nucleases were inactivated with EDTA 0.5M (PROLABO) and an incubation at 75ºC for 10 min.

Viral nucleic acids that were protected within viral particles were then extracted using the QIAamp® Viral RNA Mini Kit (Qiagen), following the manufacturer's instructions to get 40 µl of mixed viral DNA and RNA which was used for subsequent viral DNA amplification.

From here, further procedures were carried out for the extraction of the RNA fraction using a fraction of the eluted viral DNA/RNA extraction. For each sample, a volume of 9.0 µl was treated with RQ1 RNase-free DNase (1 U/µl) (Promega) at 37ºC for 30 min to eliminate all DNA while preserving the RNA. The enzyme was inactivated with µl of ethylene glycol-bis(β-aminoethyl ether)-N,N,N',N'-tetraacetic acid (EGTA) and incubation at 65ºC for 10 min. The resulting volume contains the RNA template for retrotranscription.

Bacterial extraction was carried out for the stored pellets using the QIAamp DNA Stool Mini Kit (Qiagen) following the manufacturer’s instructions without the InhibitEX tablets.

Sequence amplification of the extracted nucleic acids

Different procedures were employed for each fraction. The RNA samples were enriched by using a Sequence-independent, single-primer amplification protocol (SISPA, Figure 40) as published by Reyes and Kim (*307*). Briefly, set of 32 primers were specially designed to have random hexamers (Nmers) at the 3’-end to randomly bind within the nucleic acids during retrotranscription instead of the regular Nmers that are normally used. The 5’-end of the primers were designed with different 20-nt sequences (resulting in the conformation: 5’ 20nt-barcode-NNNNNN 3’) to serve as labels for bioinformatic identification of the sequences as one of the benefits of SISPA is the possibility of pooling different samples for joint sequencing. The other advantage is that retrotranscription with these primers adds fixed known sequences to the cDNA fragments, providing an anchoring point for primers to start the subsequent amplification. This other set of primers were designed based on the 20-nt sequence (without the random hexamer) so that PCR can amplify any of the sequences created. A set of 30+30 (with and without the Nmers) different primers were designed (Table 9). They were checked for the formation of hairpins or homodimers with OligoAnalyzer3.1 (Integrated DNA Technologies, Inc).

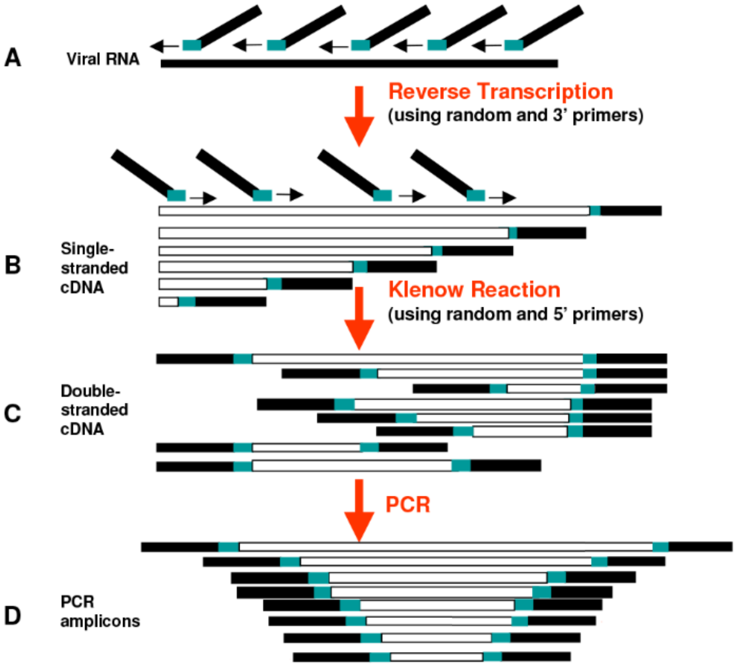


Figure 40. Outline of the SISPA method. A) Viral RNA is converted to cDNA using random-barcoded primers, labelling sequences in their 5'-end. B) Second strand DNA is synthesized using Klenow exo-DNA polymerase, in the presence of the same random-barcoded primers producing fragments labelled in both ends. C) Double stranded DNA is amplified by PCR using corresponding primer matching the barcode but with no random sequence. D) Amplicons are separated by electrophoresis and products ranging from 200–500 nucleotides. Adapted from: Djikeng *et al.*,2008 (*308*).

Table 9. SISPA primers. A set of thirty 20nt-barcodes used for amplification are shown. Another similar set containing trailing -NNNNNN 3’ was used during retrotranscription.

Name	Barcode	Name	Barcode
M1	ATCGGCTATAACTGATGCTA	M18	ACAAGCAACGGACATCTAGC
M2	AGAACTAGACCTGATCCTAG	M19	GAAGAAAACTAGTCCCCATT
M3	CGGTTAGCATTGGCGATACC	M20	TTAGTAGCCAGGATTCACGA
M4	GATCATCTCCGACTAGAGGT	M21	GAGGTCCAGTAATTCAGATA
M5	GCTAAGCGTGGTCACCTCAC	M22	CGAACATTGGTCATGGTGCT
M6	CTAGGCGAGACCAGGTAGTT	M23	TCGGAGTTTAGGAAAAGGTA
M7	TCCTTATCCTAGACGAGAGG	M24	TGTTCGTGCTAGTTTACAGA
M8	TAGCCGATGGTATTCTCTCA	M25	TCAGATTACCCAATAGTGAT
M9	ATCCAAAGGCTTCACGAGGC	M26	AGAATGCGAGTTGAGCCAAG
M10	CACACTCGCTTATGGTCACA	M27	AGACTCGGGTGTATATAGGA
M11	GGATACTTATACGTTCCCAA	M28	ACGCCTAAACCAAGTTCTCT
M12	TTACCGTGGAAGAGAGAACG	M29	CGAACAACATAGAGTGTCAC
M15	AGGAAGATCGCTCAATCGTG	M30	CTACCATCGCCAGTCCACTA
M16	GCCTTTGCCACTAAGACATT	M31	GATCTTTTCGTCTCTGTCCT
M17	CGTACAGAATCTATGACCCC	M32	GCGAGTGTGTGCTTGAGACG

Consequently, for retrotranscription, 5 µl of the DNase-treated nucleic acids (viral RNA extraction) were used for the synthesis of first-strand cDNA using the SuperScript® III First-Strand Synthesis System for RT-PCR (Invitrogen), following manufacturer’s instructions but using 50 pmol of the SISPA primers with the 6 Nmers at the 3’-end, each with a different barcode (primers M1-M10 were repeated as there were only 30 primers). Samples were then incubated at 25ºC for 10 min, followed by 50ºC for 60 min, and 80ºC for 5 min to synthesize the first cDNA strand. A volume of 0.5 µl of RNaseH was added to remove the RNA template by incubating for 37ºC 20 min followed inactivation by heating at 94ºC for 3 min. The complementary cDNA strand was synthesized by adding a mix 1.5 µl of 10X Klenow buffer (New England Biolabs), 5 U of Klenow fragment polymerase (New England Biolabs) and 12 µl nuclease-free water (Invitrogen) for a final volume of 34.5 µl. Incubations were carried at 37ºC for 60 min, followed by inactivation at 75ºC for 10 min.

The cDNA sequences used for regular PCR amplification using the SISPA primers (no Nmers). Details are as follows: A 40-µl PCR mix was prepared using a Biomix PCR kit (Bioline) containing each barcode primer. Each reaction contained 20µl of Biomix®, 17 µl of H2O, 2 µl of ds cDNA, and 1µl of the corresponding 20 µM SISPA primer. SISPA-PCR conditions were 95°C for 2 min; followed by 40 cycles of 94°C for 30 s, 65°C for 30 s, and 72°C for 2 min 30 s and a final elongation step at 72°C for 10 min. A total of five replicates of the SISPA-PCR products per sample were pooled. Confirmation of correct amplification was done by 1.4 % gel electrophoresis and pools were cleaned with using NuceloFast® 96 PCR plates (Macherey-Nagel), following the manufacturer’s protocol. Purified products were run in a second 1.4% gel electrophoresis and bands between 250-500 bp were excised and purified with the High Pure PCR Product Purification Kit (Roche) following manufacturer’s instructions for gel purification.

A whole genome amplification (WGA) was carried out with the extracted viral DNA using a multiple displacement amplification (MDA) approach which is a non-PCR technique using random hexamers and a high fidelity Φ29 DNA polymerase (*309*). This was carried out using the Illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare) following the manufacturer's instructions. The incubation time, set at 30ºC was variable and it was followed by the Φ29 DNA polymerase inactivation for 10 min at 65ºC. Starting at 1h 30 min, fluorimetrical measurements were done every 15 min using Quant-iT™ PicoGreen ® dsDNA Reagent to avoid overamplification. Incubation were stopped when over 40ng had been obtained. Purification of all products was carried out by adding an equal volume of sterile water to each tube, 4 µl of sodium acetate/EDTA buffer and vortexing. One hundred µl of 100% EtOH was added and tubes were centrifuged at 13,400 g. Supernatants were removed and pellets were washed with 70% EtOH, following a new centrifuge step at 13,400 g for 1 min. Supernatant were removed and DNA were air-dried for 15 min and resuspended in 20 µl TE. A 1.4% gel electrophoresis was used to confirm the existence of a smear expected for genomic samples.

For the bacterial samples, the hypervariable regions V1-3, spanning approximately the first >500 bp of the 16S rRNA gene, were amplified via PCR with universal bacterial primers based on the primers: forward E8F (5'-AGAGTTTGATCMTGGCTCAG-3') and reverse B530R (5'-CCGCGGCKGCTGGCAC-3'). Twenty-five primers forward primers were designed by tagging the E8F sequence with an additional 10 or 11-nucleotide in their 5'-end, called Multiplex identifiers (MIDs), that serves as a label for subsequent sequence identification, as the samples can be sequenced in the same pool. A complete list of MID sequences is found in Table 10.

Table 10. MID sequences used for demultiplexing (sample splitting) 454 pyrosequencing datasets. Each is attached to the E8F forward primer.

MID	Sequence	MID	Sequence
MID1	ACACGACGACT	MID14	AGTACGAGAGT
MID2	ACACGTAGTAT	MID15	AGTACTACTAT
MID3	ACACTACTCGT	MID16	AGTAGACGTCT
MID4	ACGACACGTAT	MID17	AGTCGTACACT
MID5	ACGAGTAGACT	MID18	AGTGTAGTAGT
MID6	ACGCGTCTAGT	MID19	ATAGTATACGT
MID7	ACGTACACACT	MID20	CAGTACGTACT
MID8	ACGTACTGTGT	MID21	CGACGACGCGT
MID9	ACGTAGATCGT	MID22	CGACGAGTACT
MID10	ACTACGTCTCT	MID23	CGATACTACGT
MID11	ACTATACGAGT	MID24	CGTACGTCGAT
MID12	ACTCGCGTCGT	MID25	CTACTCGTAGT
MID13	AGACTCGACGT		

For the amplification, a 40-µl PCR reaction was prepared using a Biomix PCR kit (Bioline) containing each barcode-forward-primer set. Each reaction had 20µl of Biomix®, 17µl of H2O, 1µl DNA, and 1µl of forward and reverse primer. PCR conditions were 95°C for 2 min; followed by 25 cycles of 95°C for 30 s, 55°C for 1 min, and 72°C for 1 min; and a final elongation step at 72°C for 10 min. PCR products were confirmed by gel electrophoresis on a 1.4% agarose gel and purified by ultrafiltration using NucleoFast® 96 PCR plates (Macherey-Nagel) according to the manufacturer's instructions.

High-throughput sequencing

Each set of amplified products was processed differently.

The bacterial DNA amplicons for 16S profiling were quantified with Quant-iT™ PicoGreen ® dsDNA Reagent (Invitrogen) and combined in equimolar ratios (200 ng from each sample) in pools containing 10 or 11 samples. Adaptors were ligated during library construction. These constructs were processed in a 454 pyrosequencing Genome Sequencer FLX with Titanium plus reagents (Roche) by the Sequencing Service from the Genomic and Health, FISABIO, in Valencia (Spain), following the manufacturer's instructions. Each pool was sequenced on an eighth of a PicoTiterPlate device.

Purified viral DNA WGAs were quantified with Quant-iT™ PicoGreen ® dsDNA Reagent (Invitrogen) and were diluted or concentrated using a Savant SPD111V SpeedVac (Thermo Scientific) depending on the result accordingly, to a final concentration of 0.2ng in 5µl, measured in a Qubit Fluorometric Quantitation (Thermo scientific). Libraries were prepared with Nextera XT kit (Illumina) using dual indexes (Table 11) following the manufacturer's protocol.

Table 11. Dual index used for Illumina sequencing of the viral DNA samples.

Index 1	i7 Bases	Index 2	i5 Bases
N701	TAAGGCGA	S501	GCGATCTA
N702	CGTACTAG	S502	ATAGAGAG
N703	AGGCAGAA	S503	AGAGGATA
N704	TCCTGAGC	S504	TCTACTCT
N705	GGACTCCT		
N706	TAGGCATG		
N708	CAGAGAGG		
N709	GCTACGCT		
N710	CGAGGCTG		
N711	AAGAGGCA		
N712	GTAGAGGA		

Sequencing was carried out in an Illumina MiSeq platform with MiSeq Reagent Kit v3 (Illumina) following the manufacturer's instructions for Paired-End 2x300 sequencing in the Service for Massive Sequencing NGS, Sistemas Genómicos, Paterna (Spain).

For the viral cDNA, total DNA was quantified with Quant-iT™ PicoGreen ® dsDNA Reagent (Invitrogen) and combined in equimolar ratios in pools of four samples each. Fragments of 200-500 bp were purified with Agentcourt AMPure XP magnetic beads (Beckman Coulter) according to the manufacturer’s protocol for that size range. The resulting fragments were quantified in a Qubit 2.0 Fluorometer (Thermo Scientific) and size was confirmed with an Agilent DNA 1000 microfluidic chip (Agilent Technologies).

Libraries were prepared using NEBNext Ultra DNA Library Prep Kit for Illumina (New England Biolabs) following the manufacturer’s protocols for small quantities of DNA using NEBNext Index 1–12 Primers for Illumina (New England Biolabs).

NEBNext Index	**Sequence**
1	CGTGAT
2	ACATCG
3	GCCTAA
4	TGGTCA
5	CACTGT
6	ATTGGC
7	GATCTG
8	TCAAGT
9	CTGATC
10	AAGCTA

Sequencing was carried out in an Illumina MiSeq platform with MiSeq Reagent Kit v2 (Illumina) by the Sequencing Service from the Genomic and Health, FISABIO, in Valencia (Spain). following the manufacturer’s instructions for Paired-End 2x200 sequencing.

B) BIOINFORMATIC PROCEDURES

Sequences from the three datasets generated by high-throughput sequencing (16S rDNA profiles, the viral DNA set and the viral RNA set) were different due to the methods employed. The objective in all three was to obtain a contingency table bearing abundance and a taxonomic information. The processing steps carried out for these data vary as follows:

Bacterial 16S profiles

Sequence extraction

All analyses were carried out in Kubuntu Linux 14.04. The regular output of the 454 platform, standard flowgram files (in binary .sff format), were generated using the 454 Genome Sequencer FLX System Software Package 2.3 (Roche) for filtering low quality reads and ambiguous assignations (<30 out of 40 Phred quality score). Sequence files were extracted with the included sffinfo algorithm (argument -s for fasta sequence file format and -q for quality qual format) for each eighth of a PicoTiterPlate. A concatenated (joined) dataset was created by including all sequences from the plate segments in a single file.

Quality processing and decontamination

The 454 platform produces long reads (>500 bp) with variable sequence sizes bearing a barcode, or MID, in their 5'-end (Figure 41). As sequencing progresses, quality drops as variation accumulates in clusters (consensus from sequences attached to a single bead is not achieved towards the end of the sequences). Homopolymers are the most common type of errors that arise during pyrosequencing but others include insertions and deletions from the PCR amplification (*310*).

Several python scripts from the pipeline Quantitative Insights Into Microbial Ecology 1.8 (QIIME, pronounced chime according to the authors) (*311*) were used independently for obtaining the 16S rDNA profiles. Fastq files (format including sequences and qualities) were created from each fasta and qual pair using the QIIME script convert_fastaqual_fastq.py. The PRINSEQ v0.20.4 script was used for most quality-related processing (*312*). With it, quality graphs were created for evaluating statistics of the 454 execution (prinseq-lite -graph_data and prinseq-graphs).

Sequence quality trimming (removal of low quality ends) was carried out with PRINSEQ using a five-nucleotide sliding window (one nt at a time) at the 5'-end considering a minimum mean quality of 30 (-trim_qual_right 30 -trim_qual_type mean -trim_qual_window 5 -trim_qual_step 1). Sequences shorter than 100 nt were filtered out (-min_len 100) and outputted as fasta+qual files. Also, sequences not having at least 70% entropy (low complexity reads) were removed (-lc_method entropy -lc_threshold 70).

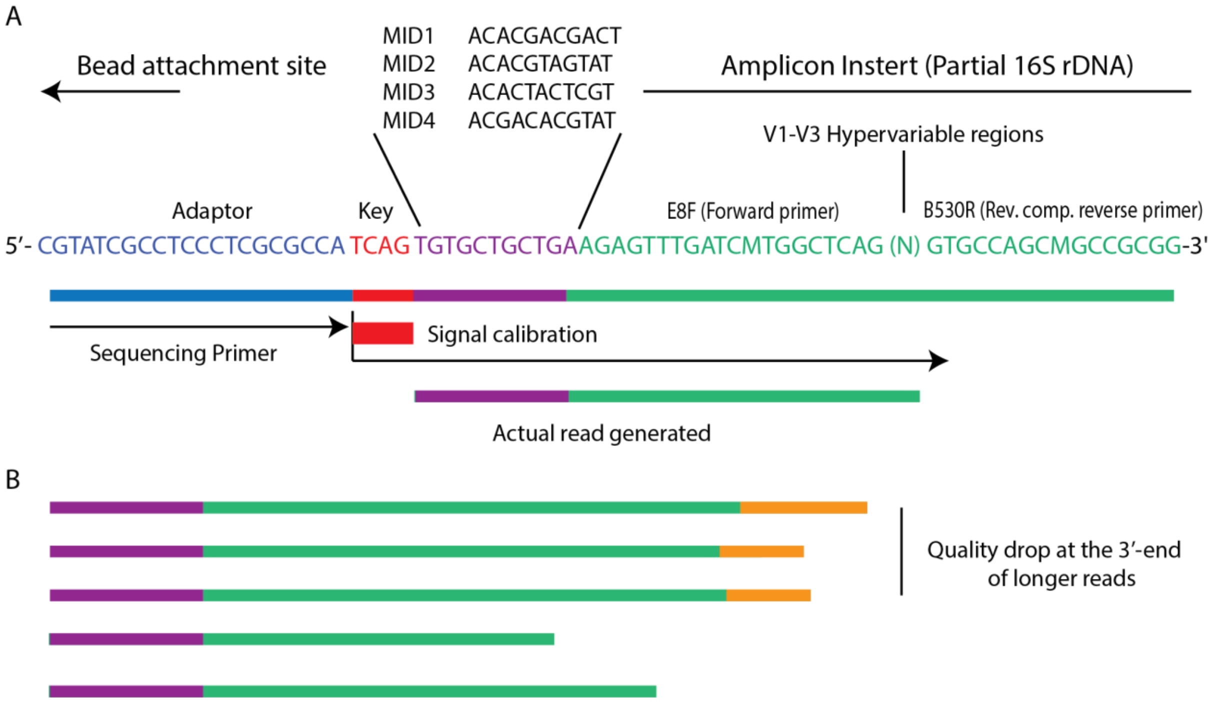

A
Bead attachment site
MID1 ACACGACGACT
MID2 ACACGTAGTAT
MID3 ACACTACTCGT
MID4 ACGACACGTAT
Amplicon Instert (Partial 16S rDNA)
V1-V3 Hypervariable regions
Adaptor
Key
E8F (Forward primer)
B530R (Rev. comp. reverse primer)
5'- CGTATCGCCTCCCTCGCGCCA TCAG TGTGCTGCTGAAGAGTTTGATCMTGGCTCAG (N) GTGCCAGCMGCCGCGG-3'
Sequencing Primer
Signal calibration
Actual read generated
B
Quality drop at the 3'-end of longer reads

Figure 41. Amplicon construct for this study. A) The structure of a 454 template sequence is shown. The sequencing primer matches the adaptor-attached sequence for elongation start at the key sequence, four nucleotides that are used for apparatus calibration (not in the output). The actual output starts at the variable MID, which is used for sample identification (purple) that is added during PCR amplification. The target sequence (turquoise) was amplified using the E8F and B530 primers. B) Output reads have variable lengths and quality drops as sequencing progresses as seen in the largest reads (orange).

A map file was manually created for usage with QIIME, which is a required base file. This consists of a table containing at least the following columns: sample id (name), barcode sequence (MID), primer sequence, group, and optional descriptions or additional groups in other columns. QIIME's split_library.py algorithm was used for splitting the different samples from each of the pools using the list of primers using the cleaned fasta+qual files for each pool and the map file for the whole set. For the splitting step, barcode was set to 11 nt (to be removed), sequence range was set to 160-800 nt (the primer was not removed as it is part of the 16S rDNA sequence), allowing 1 N (ambiguous base) per sequence, 8 nt homopolymers, up to 5 mismatches in primers (out of 20 nt as the probability of not being a 16S sequence is still low with 15 nt). The outputs are library fasta+qual files containing headers labelled according to the samples each set belongs to. These were concatenated into a single library file and a fastq was created.

DECONSEQ v0.4.3 (*313*), a script for fast identification and removal of sequence contamination from metagenomic datasets, was used for removing human datasets. The algorithm uses the Burrows-Wheeler Aligner (BWA) to search a database (*314*). In this case, decontamination was carried out in the libraries fastq file using the GRCh38 human genome (assembly published on Dec, 2013). Sequences sharing 95% identity with the human references, spanning at least 90% of the query length, were removed. The resulting sets were considered human-free and were used for the rest of the study.

One of the most important steps in 16S profiling is the formation of groups of sequences for taxonomy assignation. QIIME scripts were run individually for creating these operational taxonomy units (OTUs), sequences grouped by shared percentage of identity. For this study, OTU identity was set to 97%, which for hypervariable regions V1-3 of the 16 rRNA gene allows an acceptable separation of most bacterial genera. Only the most important and non-default parameters are specified.

Clustering was carried out with USEARCH v5.2.236 (*315*) using QIIME's pick_otus script enabled for chimera removal with (-m usearch_ref for usearch_qf algorithm). Briefly, this quality algorithm clusters sequences bearing a fixed identity percentage and then checks for chimera formation both *de novo* and using a reference database of representative high-quality sequences (gold.fa, ChimeraSlayer reference from the Broad Institute version microbiomeutil-r20110519). Clustering was done with a reference-based approach, where reference sequences (Greengenes 97% identity clusters from Aug, 2013 (*316*)) are used for seeding new clusters. *De novo* clusters were created from sequences not matching a reference. Only OTUs with at least two sequences were kept. Heuristic searches during cluster assignation by USEARCH were set for minimizing the total number of clusters that was formed: ten maximum tries on existing clusters (--max_accepts 10), and a maximum of 256 rejects new seeds (--max_rejects 256). Representative sequences were selected with QIIME's script pick_rep_set.py set to the most abundant (-m most_abundant).

The taxonomic assignment of the OTUs was carried out with QIIME's assign_taxonomy.py script using uclust search (default since version 1.8), setting maximum accepts to ten to better resolve taxonomy (--uclust_max_accepts 10) against the Greengenes 99% identity clusters from Aug, 2013. This data was used for the construction of a contingency table in the standardized BIOM binary format (*317*) using the script make_otu_table.py with default parameters. The table contains the collated information for the taxonomic assignments and is the base file for the subsequent analyses.

For phylogenetic distance measures of diversity, a tree is required. Thus, phylogeny inference for the topology was obtained from multiple sequence alignments containing all representative sequences from OTUs. The alignment was created with PyNAST v1.2.2 (*318*) using QIIME's align_seqs.py script with minimum alignment length of 150 nt (--min_length) and 60% minimum identity (--min_percent_id 0.6). This type of alignment is indicated for high-throughput datasets and works through the assignment of a set to an existing 16 rDNA alignment, in this case using a set aligned Greengenes 97% cluster representative sequences from the Aug, 2013 release as a reference. Alignments were filtered with the script filter_alignment.py (masking of unique sequences and positions containing only gaps) and FastTree 2.1.3 (*319*) was used for creating an approximated maximum likelihood tree (a scalable method for high-throughput data) using QIIME's make_phylogeny.py script.

Viral DNA bioinformatic pre-processing

The viral DNA set contained 2x300 paired-end (PE) sequences, meaning the resulting sequence for each fragment was generated in two separate reads from complementary strands that ideally overlap at their 3'-end for them to be joined bioinformatically. Hereafter, PE1 (or Read 1, R1) will refer to the 5'-3' end located at the initial part of the sequence whereas PE2 (Read 2, R2) will refer to the 5'-3' end of the complementary strand, located at the terminal part of the same fragment. The initial steps were carried out in a Kubuntu linux 14.04 PC and most scripts were written specifically for this purpose.

To understand the initial steps of the bioinformatic methodology used for the Illumina sets it is important to define the library structure (Figure 42A), the order, and some terms from the general workflow of this sequencing technology. The Illumina MiSeq platform separates sequences automatically as part of its normal workflow in a very reliable process, as the indexes are read in separate sequencing phases (Figure 42B). There are four sequencing phases in PE-sequencing, producing PE1, index 1, index 2, and PE2 in that particular order, each using a different sequencing primer to start in the correct position. PE1 and PE2 always have an equal size (sequencing-reagent-dependent). Index 2 may or may not be present, depending on the total number of samples that may be sequenced together because permutations of the two indexes allow for the sequencing of up to 96 of them in a single run but one index suffices for few samples.

The size of the fragment that is sequenced to produce PE reads is known as insert size and is one of the most important factors affecting Illumina sequencing (Figure 42C). The whole process works well with fragments spanning 350-550 bp. Fragments longer than this may have little or no overlap in their resulting reads, making them impossible to join to form a single sequence. On the other hand, shorter sequences usually get partial or complete sequences form the primer, index and adaptor flanking the insert’s 5’ end (they are not normally sequenced), followed by a poly(A) and next random nucleotides that are inputted when no template is found. The run may stop prematurely when a large proportion of the fragments are like this.

Quality statistics for the raw sequence sets were obtained with FastQC v 0.11.3 (*320*) and PRINSEQ (prinseq-lite -graph_data and prinseq-graphs). Dereplication (removal of repeated sequences) was carried out with the PRINSEQ algorithm by removing both the exact duplicates and the reverse complement exact duplicates (-derep 14).

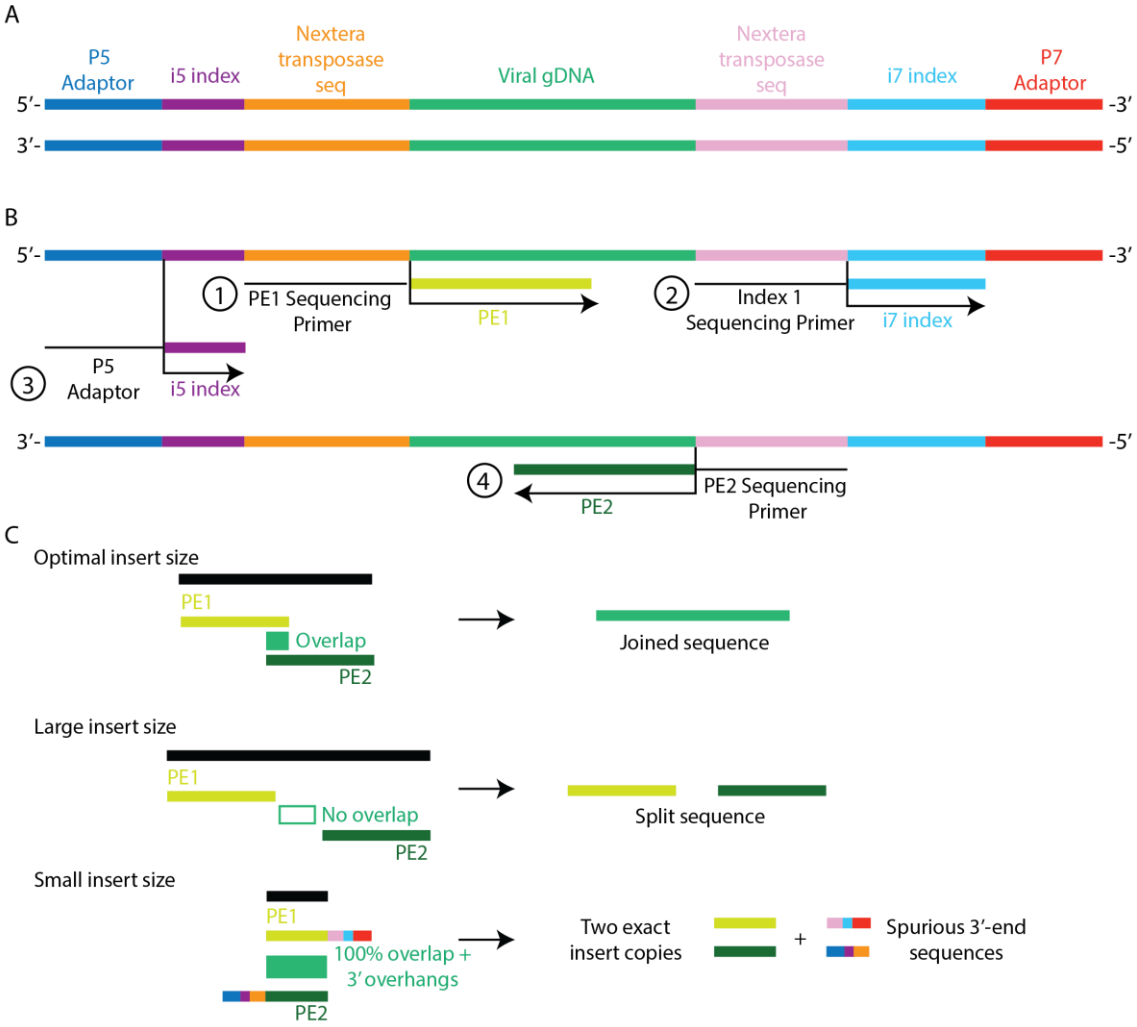


Figure 42. Illumina Nextera XT construct structure for viral DNA with sequencing phases and inserts. A) Transposases generating the target fragments introduce flanking sequences in both sides of the insert that are used for sequence primer anchorage. Dual indexes and adaptor sequences are incorporated during library preparation. B) Sequencing is carried out in four phases, each using a different sequencing primer. Only PE2 is sequenced from the complementary strand. C) Overlapping PE1 and PE2 can be bioinformatically joined to reconstruct the sequence of an original insert. Large fragments cannot be joined if overlap is not detected. Small insert size results in the sequencing of the 3' end part of the library constructs in both strands and two exactly complementary copies of the insert.

Recovering additional reads from undetermined pools

For the viral DNA samples, permutations of dual-indexes were generated during library construction. Samples were automatically classified by MiSeq Reporter 2.4 (Illumina), the sequencer platform official software, during phases two and three of the sequencing process and written in fastq format (two sequence sets for each sample, containing PE1, and PE2 respectively, whereas no file is generated during index sequencing). Many of the sequences, however, had a small insert size meaning many of them had spurious sequences from the library constructs in their 3’-ends, sometimes including the indexes. This information was used to create a method for recovering additional reads sequences that could not be separated due to faulty index reading (contained in pools called Undetermined R1 and R2 in a regular Illumina run) and assigning them to existing samples (henceforth, sequence scavenging). The method can also be adapted for index identification only (e.g. many sequencing services do not disclose the indexes that are used for sequencing).

In order to determine the construct structure in the sequenced dataset, adaptors and primers were compiled from the official sequence lists available in the Illumina website (*321*) in a fasta file that contains sequences for all types of libraries, including Nextera's and was required to identify the exact sequences and their orientations in problematic sequences. All potential adaptors and primers were searched in the largest sample's PE1 and PE2 files in forward and reverse orientations. From the complete list, four files were obtained: Fwd_adaptors_in_R1.txt, Fwd_adaptors_in_R2.txt Rev_adaptors_in_R1.txt and Rev_adaptors_in_R2 and the exact library construct structure was manually determined from these.

Once adaptors were identified, the eight-nucleotide indexes were expected to be just downstream, making them easy to identify and isolate. This information was used to detect which indexes were employed in the viral DNA Illumina run and execute a home-made shell script named Scavenge_reads_with_no_index.sh to recover some previously unidentified sequences from the "Undetermined" PE1 and PE2 pools (those having all sequences that could not be identified as belonging to existing samples). This process was defined as scavenging. It is possible only when reads are short enough for the 3' end adaptor structure to be sequenced and uses Cutadapt 1.9.1 (*322*), a program normally used for trimming adaptors. Briefly, the script scans the 3' ends of undetermined sequences for the part of the Nextera constructs that contains the index, allowing for up to 1 mismatch. It is done in two steps, a first one for the R2 indexes (i5) and a second one for each of the resulting new sets using the R1 index (i7). During scavenging, target indexes with a partial transposase sequence (7 nt) attached to the 5'-end were searched in the 3' end of the sequences. In both steps the adaptor+index match overlap was required to be complete (15 nt long), one mismatch was allowed and sequences shorter than 77 were eliminated (27 nucleotides of Nextera transposase sequences were expected to be present upstream so 77 would produce 50 nt worth of useful sequence). Any nucleotides downstream were removed. Only those pairs having both indexes and passing the size threshold were included in their original sample pools as additional scavenged reads.

Quality processing

Cutadapt was used to remove partial adaptors sequences (≥13 nt) that had passed full length sequence adaptor filters. Adaptor fragments were selected from the library construct structure that was previously determined and were set to survey both the 5' and 3' flanking regions on the PE1 and PE2 reads (parameters -a, -g, -A, and -G), with an overlap of twelve nt, and error tolerance of 13% (-O 12 -e 0.13).

Quality control was carried out with PRINSEQ. Sequences with an overall mean quality lower than 25 were removed (-min_qual_mean 25). A five-nucleotide sliding window (advance step of three nt) was used to remove segments with average quality lower than 33 (-trim_qual_right 33 -trim_qual_type mean -trim_qual_rule lt -trim_qual_window 5 -trim_qual_step 3), plus a fixed trim of 4 nt was applied in both the 3' and the 5' ends (-trim_tail_left 4 -trim_tail_right 4). Minimum length was set to 100 nt and a low complexity filter was applied with entropy threshold of 72% (-min_len 100 -lc_method entropy -lc_threshold 72). An additional dereplication step was also included as the probability of two sequences being exactly the same in a WGS library is close to none with this sequencing depth. Both forward and reverse exact matches were considered (-derep 14). When using Paired-end datasets, PRINSEQ produces four files as output sequences passing the filters: paired-end sets and two singleton files, one for PE1 and one for PE2. This file contains reads passing filters in only one of the PE sets (unpaired reads). As no further processing is required for these, singletons were joined into a single file and changed to fasta file format (removing quality information).

Join paired sequences

Most sequence joiners do note cope well with short reads. As a general feature, they assume that the resulting read must always be longer than any of the paired sequences, thus expecting the user to provide a large minimum overlap value. Also, some do not accept pairs in which reads have different lengths (e.g. those that have been trimmed). Due to the characteristics of this particular dataset, COPE v1.2.5 (*323*) was use in standard overlapping mode (-m 0). The error was set to .25% and the overlap was set to be between 10 and 300 (-c 0.75 -l 10 -u 300). Both the reads that were joined and the ones that remained split were converted to fasta format. Joined reads and singletons from the cleaning step were concatenated (JnS sets). Sequences that could not be joined were kept in two other separate files. These three files contain the sequences that were used for the rest of the study: an unpaired set and two paired end set.

Viral cDNA bioinformatic pre-processing

RNA-derived viral sequences underwent additional processing steps than their DNA counterparts as the SISPA approach added inner barcodes during retrotranscription of the original sequences that enabled multiplexing (pooling several barcoded samples into a single library construction). This barcode represents an artificial DNA sequence that must be ultimately removed. The set was sequenced using Paired-End 2x200 reads (Figure 43) and bioinformatic effort was focused on recovering the maximum number of sequences. The short fragment size deemed Nextera unusable as transposase cut sequences into shorter still segments. Libraries were instead constructed using NEBNext Ultra DNA Library Prep Kit for Illumina (New England Biolabs), which in turn is based on the TruSeq libraries (Illumina), with only one set of primers (single index): NEBNext Index 1–12 Primers for Illumina (New England Biolabs) on the 3’ end of the forward strand (and complementary sequence). This type of library ligates adaptor-containing sequences to both ends of the fragments (Figure 43), while retaining the original size of the insert.

Sample separation was carried out with the MiSeq Reporter 2.4 (Illumina) and fastq files were generated for all files. Raw sequences were examined with FastQC v 0.11.3 (*320*) and PRINSEQ (prinseq-lite -graph_data and prinseq-graphs) to get the initial run statistics. Several preprocessing steps were identical between the DNA and cDNA sets, such as the sequence dereplication, which was achieved with PRINSEQ following the same strategy as for the viral DNA sets. Only differences and key details will be described below. It is important to note the “samples” that were separated by the MiSeq apparatus were in fact pools containing multiplexed reads.

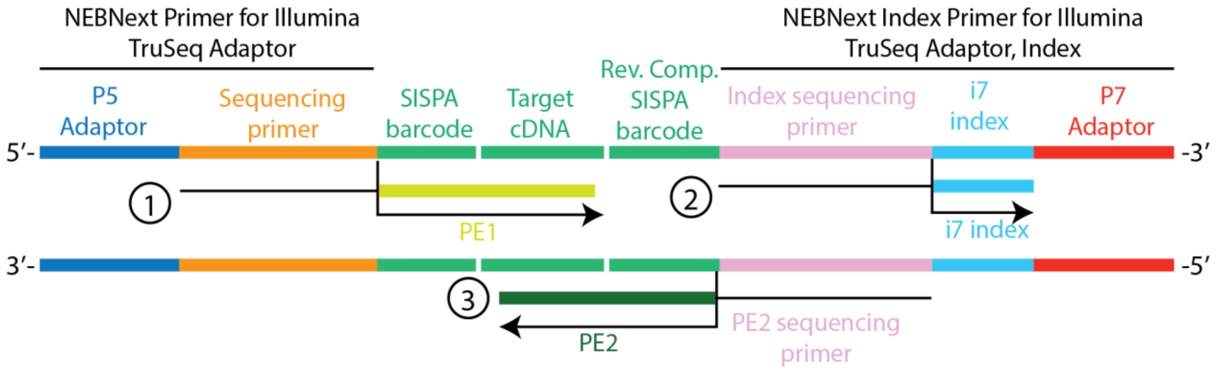


Figure 43. NEBNext library construct for cDNA datasets. These sequences are equivalent to the Illumina TruSeq. Only one adaptor was included (i7 index). The PE1 and the index are sequenced in the forward strand and the PE2 is sequenced from the complementary strand in that order. Additionally, the cDNA insert contains SISPA primers at its 5' and 3'-ends, enabling pools of multiplexed samples. The barcode is the same in both strands and normally only the 5'-end SISPA barcode is sequenced with regular insert size.

Scavenging reads from undetermined pools

Due to the use of SISPA primers, sequences were pooled in ten different libraries in the cDNA set, each having a different index located after the sequence in the 3'-end, apart from the internal SISPA barcode which was considered part of the insert. In sequences having a short insert size, the adaptor in the PE1 was detected as it was attached at the 3'-end, followed by a poly(A) tail and random noise. In these cases, the index could be recovered as part of the read. The exact structure of the construct was determined by analyzing the 3' ends of the short insert reads in the same way the construct structure for the viralDNA set was obtained. Briefly, by searching official adaptors, (since NEBNext adaptors are equivalent to TruSeq's) in a representative dataset, existing primers and adaptors, as well as their correct orientations, were identified. Each of this and their positions were used to identify the library construct structure.

Once this was determined, a scavenging process similar to that of the viral DNA set was carried out, using a simpler version of the Scavenge_reads_with_no_index.sh script that only searches for adaptors and index in the 3'-end. Briefly, the scavenger uses Cutadapt v1.9.1 to scan the Undetermined sequences pools PE1 and PE2 (containing sequences that could not be separated due to faulty index reading) to detect a list of TruSeq adaptor plus the index. In this case, the overlap was set to 39 and up to 0.05% error. The recovered paired reads were concatenated with the pools they belonged to.

In the rare event of detecting an index in the 3'-end of a sequence not belonging to the correct library (e.g. index 3 found in a set that should only have index 1), supported by a high-quality score (> 30), these were filtered using the home-made script, remove_spurious_indices.sh. This script is based on the scavenging script that analyses all the spurious indexes and removes the sequences from both PE files. Those containing a single mismatch in the detected adaptor+index sequence were also removed. Cutadapt was used to remove the other regular adaptors (those expected to be generated as part of the normal construct) from both PE1 and PE2 in their 3’ and 5’ ends (passed with arguments -a, -g, -A, and -G) with a minimum overlap of 18 nt and a maximum error rate of 0.15% (-O 18 -e 0.15).

Separating samples

Adaptor-free sequences were used for demultiplexing the pools into separate samples. To do this, sequences are filtered according to the different internal barcodes that were introduced during reverse transcription as part of the SISPA protocol. However, different problems must be dealt with. Due to the way libraries were constructed, only the four barcodes that were supposed to be present in each library (pool) were used. To process all samples, a suite of four home-made perl scripts was created to scavenge sequences containing incomplete or internal SISPA barcodes, executed with a pipeline named Illumina_SISPA_scavenger_pipeline.sh.

In order of execution, the first step was executed by the home-made script Illumina_paired_ends_detect_multibarcode.pl which managed the removal of sequences that had more than one barcode (this artifact is relatively rare with SISPA barcodes, and is seen as tandem barcodes). The script cut sequences from the 5' end in a fastq file whenever more than one barcode is found to keep just the one downstream to make sure each sequence started with a barcode. The sequences bearing multiple barcodes were eliminated due to their small size after barcode removal. Since the SISPA barcodes were 20 nt long, variation was allowed at their 5’-end (in their first 5 nt). Using the remaining sequences, the pipeline then performed a demultiplexing step using the exact barcodes only. This was carried out by using the fastx_barcode_splitter.pl script from the FASTX-Toolkit v0.0.13 (*324*) for barcodes in the 5’-end (-bol) and no mismatches allowed (--mismatches 0). Sequences bearing the same index in both PE sets were considered correct and were stored without the index sequences.

Those sequences that could not be identified were the target for a second round of demultiplexing. Prior to this, a scavenging step was included to recover potential targets bearing partial barcodes with the home-made script Illumina_paired_ends_fastq_barcode_scavenger.pl. Sequences with tandem-repeated partial barcodes were eliminated. The script searches the last ten nucleotides from each barcode in the unmatched set produced by the FASTX-Toolkit (do not confuse with Illumina’s Undetermined reads) and generates sequences starting with truncated 15 nt barcodes. A second round of sample demultiplexing was carried out with the fastx_barcode_splitter.pl script in this new set, this time with less stringent parameters by searching for 15 nt barcodes (truncated at the 5’ end) and allowing 2 mismatches (--mismatches 2). Again, sequences bearing the same barcode in both PE sets were considered for further analyses, their barcodes were removed and they were concatenated with the sequences from the first demultiplexing round, generating new cumulative PE files (a pair for each sample).

At this point, sample sets contained only sequences bearing the corresponding indexes in both sequences (PE1 and PE2 simultaneously). Two new sets were compiled containing sequences that had at least one missing barcode in each pair by concatenating those unmatched in any of the demultiplexing rounds. The home-made script Illumina_paired_ends_unmatched_cleaner.pl (part of the same pipeline), permanently discards those pairs of sequences that fail to get a barcode in both the PE1 and the PE2 sets simultaneously. The rest have at least one correct barcode of the sequences and were recovered by assigning them with their only barcode using the home-made script (the last in the pipeline) Illumina_paired_ends_adopt_orphans.

An exploration to recover unidentified sequences was carried out with Cutadapt by using the non-barcoded results from the last scavenger. The script, called Last_survey.sh searches both the 5'- and 3'-ends for the SISPA barcodes in the respective orientation) and uses Cutadapt option to demultiplex accordingly, requiring a 12 nt overlap and 0.2% error. The approach only works with one dataset at a time so the resulting R1 and R2 sets were processed with a perl script (Get_Paired_Scavenged_sequences.pl) to extract only those with matching indexes in both sequences. The main difference with previous similar operations is that this approach searchs for the reverse barcode sequences, produced when the 3'-end of the construct is sequenced due to the limited insert size.

The resulting PE reads from the four demultiplexing steps were aggregated in corresponding sample fastq files, forming the definitive sets that were used for the rest of the processing. In summary, the datasets were formed by a large majority of pairs having the exact SISPA barcodes in both strands (PE1 and PE2), pairs with matching barcodes having some mismatches in the initial nucleotides, pairs having one good barcode in a sequence but missing in the other, and pairs with matching barcodes in which one is found as part of the 3' construct in reverse orientation.

Quality processing

The removal of the remnants of the flanking SISPA primers was carried out with Cutadapt using a two-step process automated for each sample with the home-made script remove_adapter_remnants.sh. Briefly, the iteration of Cutadapt removes the complete SISPA barcode with a minimum overlap of 15 nt and 25% error (-O 15 -e 0.25) in forward orientation near the 5'-end and the reverse-complementary in the 3'-end. The second iteration filters truncated versions in the same orientations (for those that avoided detection in the first iteration due to sequence changes, especially those in the 3'-end). New quality graphs were generated with PRINSEQ ((prinseq-lite -graph_data and prinseq-graphs) to decide the trimming and filtering strategy. Trimming was carried out with PRINSEQ as well, using a three-nucleotide sliding window (three-nt step) at the 5'-end with a minimum mean quality of 30 (-trim_qual_right 30 -trim_qual_type mean -trim_qual_rule lt -trim_qual_window 3 -trim_qual_step 3). Fixed trimming was carried out to remove the first five positions in the 5'-end and the last five in the 3'-end (-trim_tail_left 5 -trim_tail_right 5). Additionally, reads shorter than 50 nt and sequences with an average quality lower than 20 were filtered-out (-min_qual_mean 20) and low complexity filters were applied to sequences with less than 70 entropy (-lc_method entropy -lc_threshold 7). Resulting singletons were converted to fasta format as no further processing was required.

Overlapping paired reads were joined with COPE v1.2.5. The error was set to 0.25% within the overlap, minimum overlap is 10 and maximum is 200 (-c 0.75 -l 10 -u 200). All samples were then converted to fasta format and joined reads were concatenated with their corresponding singletons from the cleaning step in files (JnS). Sequences that remained split were not discarded. The sequences that were used for the rest of the study were included in three files per sample: an unpaired set and two paired end sets.

Cluster formation.

Processing of large datasets is usually time-consuming and requires vast computational resources particularly during homology searches. Thus, the datasets were clustered in order to ease some parts of the taxonomic processing.

Joined reads and singletons from all samples were concatenated in a set henceforth called DNA_JnS, and sequences that could not be joined were concatenated in a file (including both the PE1 and PE2sequences) hereafter called DNA_split containing headers created to track their origins. Both sets represent the end of the preprocessing steps for viral DNA sets. An analogous procedure was carried out for the cDNA set, generating the cDNA_JnS and the cDNA_split set. All the sequences in the four sets had their sample identifications included in their heads (sequence names) so that they could be later demultiplexed again. The following steps were carried in Xeon servers and a cluster running CentOS 5.9.

Each of the four datasets was clustered using the 64-bit version of USEARCH 6.1.544 (-cluster_fast) at 99% sequence identity (-id 0.99) setting maximum accepts at two sequences and maximum rejects at 32. A map of the cluster composition was created with the .uc file (default cluster information tabular file from USEARCH) with an inhouse perl script get_cluster_composition_map.pl. This algorithm writes the total number of items in a cluster, its centroid (representative sequence) and all samples that are contained within it.

Constructing databases

In order to correctly identify the sequence taxonomy, different nucleotide databases from human, bacterial, fungal, archaeal, viral and prophage sequences were created for this study. This section details the procedures for building each.

Human Database

Since it was intended just for filtering human contamination, the human database was also the simplest one. Chromosome sequences, mitochondrial DNA, RNA transcripts, and unassembled reads were obtained from the National Center for Biotechnology Information (NCBI), contained in the Assembly GRCh38.p7 - *Homo sapiens* Annotation Release 108 from June 6th, 2016, downloaded on Aug 8th, 2016. The assembly information is available at http://www.ncbi.nlm.nih.gov/genome/annotation_euk/Homo_sapiens/108/. All sequences were concatenated in a single fasta file called HumanDB_2016_08_08.

Fungi, bacteria and archaea databases

The fungi, bacterial and archaea databases were created simultaneously and some of the initial steps for constructing them were shared among them. The fungi and acrchaea databases were created from two sources:

For fungi and archaea, gene source information can be obtained by parsing gene information directly from ASN.1 files (generic storage standard files for biological data) that can be obtained from the NCBI.

To deal with whole sets, we used ags files, containing binary versions of all genes from a dataset. These files were downloaded from NCBI at ftp://ftp.ncbi.nih.gov/gene/DATA/ASN_BINARY/Fungi/All_Fungi.ags.gz and ftp://ftp.ncbi.nih.gov/gene/DATA/ASN_BINARY/Archaea_Bacteria/ Archaea.ags.gz on August, 8th, 2016. Each ags file contains a very compact xml-like database that was parsed with an NCBI-tools package, gene2xml v 1.5 (*325*) to extract the Gene-source_src-str1 labels, containing Accession Numbers (hereafter Acc Nums, universal sequence identifiers) for each of the entries. Some records had no Acc Nums, but instead deprecated GIs (main id numbers in NCBI until replaced with Accession Number in September, 2016). To update these, they were searched in the file ftp://ftp.ncbi.nih.gov/gene/DATA/gene_history.gz and the updated GI was used to get an xml file to parse the Acc Num. This was achieved using NCBI Entrez Programming Utilities' Efetch algorithm, which is a programable set of retrieval tools that work with http protocol for remote access of sequences directly from the NCBI. It is executed from the browser of from command line via web access (https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=gene&id=GI&retmode=text&rettype=xml where the "GI" is substituted with the query and then the xml is parsed to extract the Acc Num) and parsing the output. A fasta file was then obtained from the resulting list by querying the Acc Nums using the Efetch online tool directly (https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nuccore&id=AC&retmode=text&rettype=fasta). As bacterial sequences are in the order of millions, a bioperl script was used instead, called bioperl_fetch_GenBank_seqs_from_acc_num.pl that essentially uses the same eutils' efetch tool but with batch queries extracting large data packages automatically. The temporal database was called gene_source.

For bacteria, archaea and fungi, complete genomes were downloaded on the same date. For bacteria, these were the only sequences that were downloaded. To achieve this, two tables containing the assembly projects were downloaded: ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/fungi/assembly_summary.txt and ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/archaea/assembly_summary.txt. The exact project ftp addresses were parsed from the tables and the genomes in .fna format was downloaded and concatenated into a single genome database that were in turn concatenated with the gene source sequences.

For fungi and archaea, in order to build a concatenated non-repeated dataset, sequences with repeated Acc Nums were removed, using complete genomes preferentially. Headers were standardized to NCBI's Sept, 2016 nomenclature (not having a GI) and sequences were dereplicated using PRINSEQ for exact duplicates or those where one is contained within other (-derep 235). A list of Acc Nums was generated for each database. The next step depends on taxonomic identification.

Obtaining taxonomic information

Taxonomy identifiers or taxids are required for official taxonomic labelling. Cross-reference containing complete lists of the existing Acc Nums and their related taxids were downloaded and concatenated. These are available from the NCBI at ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/accession2taxid/ and contain the accession, version, gi, and taxid. Using the accession numbers and the tables, the inhouse script subKingdom_part1_1col.pl (part of a larger home-made suite of scripts for obtaining taxonomy) was used to obtain a taxid for each accession number. Some Acc Nums had no traceable taxid in the tables, most of the time due to deprecated numbers but they were obtained with a simple Entrez request via eutils (https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nuccore&id=AC&retmode=text&rettype=native where AC is replaced by the Acc Num and the output xml is parsed for extracting the taxid).

The taxid that is related to the Acc Num always points to the last node (taxid) in the taxonomy (that may be the species or strain taxonomic level if available). To reconstruct a complete taxonomy, a node map was obtained from files contained in ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/. These included the nodes.dmp, a table that lists all the links between different level taxids, and the names.dmp file, a table including the scientific name and aliases of each node. The home-made script subKingdom_part2_no_fasta_species_taxid_to_stdout.pl was used for this. Briefly, the script loads all node relations, their scientific names and the rank they have and takes each initial taxid in the input to construct the whole taxonomy, starting at the outermost nodes and going up, level by level, until the root of the taxonomic tree is reached (all taxids connect to the root eventually). This version of the algorithm also outputs a relation of terminal taxids to taxids at the species level, which was used to build the database.

A second script called subKingdom_part2_no_fasta_taxid_to_wholetax.pl was used with the species taxids to create a table containing the complete 22-level taxonomy for each item (the maximum number of levels), effectively creating a taxonomy table containing each of the species in the databases. Most of these categories are normally empty as each species has different levels available, mostly depending on the domain type of organism. Regardless, the table was refined with the script Filter_General_case_tax_and_asign_to_headers.pl that extracts the domain, phylum, class, order, family, genus and species columns for eukaryotes and prokaryotes. The script also fills some of the empty categories with special labels indicating which is the closest category that is defined. These modifications are included for preventing the unclassified categories from being counted together.

Processing the databases

The list of species-level taxids was the input the home-made script split_fastaDB_by_taxid_noNs_fixed.pl to classify and separate all sequences by species in the bacteria, archaea, and fungi databases. By doing this, a different file was created for each species and sequences were cut where multiple Ns were found. Also, plasmids were separated in a different file if present.

The last step in the fungi and archaea databases was clustering the databases with 64-bit USEARCH v8.1.1831 (cluster_fast) at 99% identity (-id 0.99). The output files were renamed to FungiDB_2016_08_08.fasta and ArchaeaDB_2016_08_08.fasta and their taxonomy files were FungiDB_2016_08_08.taxonomy and ArchaeaDB_2016_08_08.taxonomy.

The largest database was, by far, that of bacteria so it required a different treatment to reduce its size. The selected approach was to construct pseudopangenomes, consisting on the largest genomes for each species plus additional sequences that did not match these. To do this, the largest genome was extracted using the inhouse script extract_longest.pl. Remaining sequences were split with a home-made script, split_seqs_larger_than.pl. The longest genome was formatted for blast (*306*) and a megablast search was executed for the fragments with 95% and spanning 90% of the query length (-perc_identity 95 -qcov_hsp_perc 90). Those that did not align to the largest genome were searched with blastn from the BLAST+ suite v2.3.0. Those that did not align in both iterations were kept as unique accessory sequences. Finally, a USEARCH clustering was carried out at 95% identity with cluster_fast (from v v8.1.1831). The resulting database was labelled as BacteriaDB_2016_08_08.fasta and its taxonomy file was BacteriaDB_2016_08_08.taxonomy.

Viruses database

Viruses are of special interest to this study and were thus the most important fraction to be explored. Consequently, their databases were the ones requiring most processing. Five different sequence sources were used to evaluate their differences for viral database building. All were downloaded on November 16th, 2016.

The first was the gene_source database, obtained similarly to that from archaea, fungi and bacteria, by parsing the binary file ftp://ftp.ncbi.nih.gov/gene/DATA/ASN_BINARY/Viruses/All_Viruses.ags.gz using NCIBI's gene2xml algorithm to obtain Acc Nums. With these, the script bioperl_fetch_GenBank_seqs_from_acc_num.pl was used to query batches and obtain the fasta files with efetch from E-utilities. Those sequences having a deprecated GI instead were searched in the file ftp://ftp.ncbi.nih.gov/gene/DATA/gene_history.gz for GI updates and Acc Nums were obtained by directly parsing xml outputs from E-utilities queried via http. All fasta files were concatenated into gene_source.fasta. Complete genomes were obtained by parsing the exact ftp addresses from the assembly projects file ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/viral/assembly_summary.txt and downloading the corresponding fna files. These were concatenated into a single file, refseq.fasta.

A different dataset was obtained directly from the packs ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/all.fna.tar.gz and ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/all.ffn.tar.gz. The difference with these sequences is that they are fully annotated instead of containing a single genomic sequence. Unlike fna files which generally contain also non-coding sequences, ffns only contain coding sequences. The fna sequences were used preferentially but some had only ffns. The concatenated file was named NCBI_FTP_Genomes.fasta. Another dataset was compiled from viral sequences obtained from the European Nucleotide Archive at http://www.ebi.ac.uk/genomes/virus.html. Although outdated (the last entry was on May, 2015). This one contains separate datasets for viruses, phages, archaeal viruses and viroids. All of them were obtained and concatenated into the file EBI_ENA.fasta. A dataset was obtained from ftp://ftp.ebi.ac.uk/pub/databases/fastafiles/embl_genomes/genomes/Phage/, containing only complete phage genomes. It was last updated on January, 2015 so it was outdated as well. The files were concatenated into a fasta file called EBI_Phage.fasta.

Finally, and most important, the complete collection of sequences from the International Nucleotide Sequence Database Collaboration (INSDC, including sequences from NCBI, EBI and DDBJ) was downloaded from the EBI's ftp site ftp://ftp.ebi.ac.uk/pub/databases/fastafiles/emblrelease/. Additional updates were downloaded from ftp://ftp.ebi.ac.uk/pub/databases/fastafiles/emblnew/. These are separated into groups of organisms. In the case of viruses, they are subdivided in Expressed Sequence Tags (EST), Genome Survey Sequences (GSS), High Throughput Genome sequencing (HTG), the majority located in the standard fraction (STD), and Transcriptome Shotgun Assembly (TSA). Patented sequences were ignored. All these sets and their respective updates were downloaded and concatenated into the INSD.fasta dataset.

The contents of all five datasets were first compared by looking for shared accession numbers and header names. Those that were repeated were removed and the remaining ones were concatenated into a single viral database. Headers were standardized to comply with NCBI's current naming standards (only Acc Num and name). In the end, all sequences had an Acc Num.

Similarly to the other databases, taxonomy was built with all extracted Acc Nums by using the subKingdom_part1_1col.pl script to get taxids, and those with no hits in the tables were obtained with E-utilities http queries. The script subKingdom_part2_no_fasta_species_taxid_to_stdout.pl was used to generate a list of associated taxids at the species level and the the subKingdom_part2_no_fasta_taxid_to_wholetax.pl script was used create 22-level taxonomy table, that was filtered with the Filter_General_case_tax_and_asign_to_headers.pl. This last script was modified to scan nodes with no ranks (meaning they cannot be attributed an official taxonomy rank such as genus or species) in search for the Baltimore classification. Found taxonomic levels were included in a resulting 7-level taxonomic table containing the main categories for viruses: Virus type, Baltimore classification, Order, Family, Subfamily, Genus, and species. Virus types included Bacteriophage, Archaeophage and Virus. The resulting taxonomy was stored in file VirusDB_2016_11_16.taxonomy.

The species-level list was used with the script split_fastaDB_by_taxid_noNs_fixed.pl to separate the sequences in the virus database into species by using the associated Acc Num. Each species had its own fasta, which was used for a pre-cluster step using blast searches (from the BLAST+ v2.5.0) against themselves by formatting with makeblastdb then running a megablast search with 100% identity and 90% query coverage (-perc_identity 100 -qcov_hsp_perc 90) and printing a tabular output (-m 6). By parsing the blast output, clusters were formed with the homemade pick_rep_set_from_blastout.pl script, which detects the largest sequences containing the most sequences within. Each resulting species dataset was then fragmented into 1,000 nt fragments with the split_seqs_larger_than.pl script and the largest sequence was extracted to do a pseudopangenome search similar to that in the bacterial database but using USEARCH v8.1.1831 (-cluster_fast) with 99% identity (-id 0.99). The resulting clusters were concatenated to form the database, which was called VirusDB_2016_11_16.fasta and had its complementary taxonomy, VirusDB_2016_11_16.taxonomy.

Prophages database

A prophage database was created based on sequences from the PHAge Search Tool (PHAST) database (*326*). Basically, the PHAST database is based on the results of searching viral genes in bacterial genomes (mostly). There are three criteria used by the PHAST tool to identify prophages after identity search. The first is to recognize a sequence having the same number of CDS that phage it is suppose have, yielding the max score. The second method two is to calculate phage prevalence among the possible candidates, when more than one is found and score depends on what percentage it contributes with. The third method consists in granting added score bonuses for each predicted proteins matching phage-related keywords (e.g. capsid, head, etc.), if size is longer than 30 Kb, more than 40 proteins are detected of phage compose >70% of the proteins in the region.

Since the database is automatically generated from the results of the bioinformatic tool, it has several limitations and must be curated before usage. The fasta was downloaded from http://phast.wishartlab.com/Download.html (Mar 15, 2016 update) as well as the table displaying query genomes that are used for PHAST searches and the individual reports. The details files are summarized in a single table containing information of all putative prophage regions in the genomes. Only those appearing in both the database and either table are conserved (i.e. those sequences for which fragment statistics could be obtained). To filter the sequences that may be useful, a table was downloaded, containing the statistics of all the available bacteriophages genomes, from NCBI at https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=10239&opt=Virus&sort=genome (taxid10239.tbl) the real information on the bacteriophages was obtained with R (*327*). Categories considered were genome length and number of proteins from the actual phages, and prophage region size, total CDSs, hypothetical proteins, and bacterial protein number in the region from the putative phages. PHAST sequences were filtered according to the analysis to resemble the actual phage information. The remaining sequences inherited the Acc Num of their bacterial host and were considered for subsequent analyses and the headers were standardized.

Taxonomy was obtained for their host bacteria using the same procedure as for the viral database but, after the filtering step by the Filter_General_case_tax_and_asign_to_headers.pl step, the term prophage was added to all levels and type of viruses were set to “prophages”. No clustering was carried out for this dataset. The resulting database was called PHAST_2016_11_16.fasta and had the associated PHAST_2016_11_16.taxonomy.

Taxonomic identification for viral sequences

As we published in two earlier studies regarding viral sequence assemblies (for 454 and Illumina, respectively), using raw reads is viable for taxonomic assignment as it is not as dependent on large sequences unlike the functional analyses (*298*, *328*). Yet, taxonomic assignment is not straightforward and requires robust methods for avoiding false positives.

Consequently, taxonomy assignment was achieved with different approaches that were carried out using the clustered viral datasets and the newly created databases. Four clustered datasets were created, all produced at 99% identity for temporary processing of large datasets: DNA_99clu_JnS.fasta (clustered joined and singletons from the viral DNA set), DNA_99clu_split.fasta (clusters from pairs that could not be joined from the viral DNA set), cDNA_99clu_JnS.fasta (clustered joined and singletons from the viral RNA set), and cDNA_99clu_split.fasta (clusters from pairs that could not be joined from the viral RNA set). The databases used were the ArchaeaDB, BacteriaDB, FungiDB, HumanDB, VirusDB, and PHASTDB.

In summary, the taxonomic identification was carried out using each representative sequence in the cluster files to identify homologous sequences in all the datasets. Instead of filtering out sequences other than viruses, these were used for comparison of other DNA sources that may in fact be present in the datasets. Unspecific results, *i.e.* sequences that simultaneously matched entries from different databases, were discarded as they bore no relevant taxonomic information. The rest were assigned a consensus taxonomy displaying the last common ancestor (LCA) among the potential hits within a given database.

An additional file was created for each of the datasets, containing the number of items per cluster (from the cluster map files generated before) and the length of the representative sequences. Homology searches were carried out with algorithms from the BLAST+ suite v 2.5.0 and sequence mappings using Bowtie2 v 2.2.3 (*329*). Although it may seem redundant, there is some variation among the different types of blast searches because of the intrinsic program heuristics (e.g. megablast runs faster but may ignore some 100% identity hits whereas blastn may ignore some larger sequences that are reported by megablast). Therefore, each set was surveyed using both megablast and blastn, as well as with Bowtie2. Viral databases were additionally searched with tblastx for detecting more distant homologs that were conserved at the amino acid sequence (six-frame translated query sequences versus six-frame translated database items). In summary, a total of 72 search strategy permutations (considering all databases, datasets and three algorithms) were executed out plus four specific tblastx searches with the VirusDB. This was automated and executed in two Xeon servers running CentOS 5.9.

All databases were formatted for blast and bowtie indexing (makeblastdb and bowtie2-build, respectively). The datasets were split into multiple fragments and searches were run in parallel with inhouse scripts.

The megablast and blastn searches were run with identity percentage of 80% over a minimum span of 80% of the query sequence length (-perc_identity 80 -qcov_hsp_perc 80) and results were written into summary tables additionally containing the query length, taxid, and scientific name (-outfmt '6 qacc sacc pident length mismatch gapopen qstart qend sstart send evalue bitscore qlen staxid sscinаme'). Only one alignment was reported for each resulting hit in the databases (-max_hsps 1) but the total alignments formed with different hits and the ones reported were unrestricted (both -max_target_seqs and -num_descriptions have 500 results by default). The tblastx searches were run with the same parameters except for the identity percentage, because the program does not have such parameter (it had to be filtered posteriorly).

To make the approaches comparable, bowtie mappings were carried out with global alignments (--end-to-end) with low stringency parameters (--very-sensitive) and a mode that reports all alignments (-a) instead of the usual best hit conserving the input order (--reorder). The output SAM header was ignored (--no-head) and was parsed from the stdout to create a tabular file containing the query, type of alignment (strand and orientation), hit, mapping position, and CIGAR string (summarized coded alignment).

A perl script, summarize_multiple_blast_outs_into_single_hit_table.pl, was created to aggregate results into a single table. It used a file with the length of each sequence receiving a series of thresholds for identity, raw length and query coverage percentage to filter the tables. This script takes a list of several tabular result files that were run with the same query against different databases, and then builds a table with the total number of species per search method for each query. Depending on the method (megablast/blastn, tblasx, or bowtie2), the output was pre-processed differently. For instance, the length in the tblastx results was converted to nt length as they were originally in amino acid length.

Also, bowtie results (hereafter bwt) are reported with no quality estimate when multiple hits are allowed per query. In order to cope with this limitation, the CIGAR string (sequence alignment in abbreviated format, containing no exact positon but sizes of alignments or non-aligned segments) was used by the algorithm to calculate the length, query coverage and indels. The latter was used for an approximation of sequence identity (indels and softclipped bases were counted as mismatches), which is somewhat close to the identity in blastn, albeit not perfect since such alignments do not distinguish between matches and mismatches (as a mapper, bwt only stores whether or not it is aligned in order to consider heteroplasy). To adjust this value, the calculated sequence identity was compared to actual identity percentage values from blastn results for the same query+hit pairs using inhouse R scripts. Only query-hits pairs appearing in both the blastn and the bwt results were considered (one per query). Both the query and the hit were required to match between the two methods. The matching bases were counted in the CIGAR string and divided by the total length of the query sequence. The percentage of sequence that was aligned was calculated but could not be compared directly with the percentage identities. Bwt2 hits were global and therefore had longer alignments, which influence the estimated identity percentage. This means that shorter alignments in blastn can have better identities because they drop the unaligned part that bwt2 cannot. Thus, both scores were relativized using their lengths so that the remaining difference is explained by mismatches occurring (and the differences would be the percentage of identity variation explained by mismatches). The median of the difference between the two scaled methods was subtracted to adjust this value and a format equivalent to that of blast outputs was created for the bowtie results using the homemade script Filter_BLAST_and_Bowtie2_dummy_m8_creator.pl.

The output was a table displaying each cluster, its size (total sequences it contains), and the total number of different species identified for that query by each of the search approaches selected. A thorough comparison of the result tables from the different methods was carried out using R for determining the thresholds for several parameters. Particularly, the tblastx identity cut-off was set to 57 %, whereas the rest was set to 85 %. The minimum length for all methods was set to 103 nt and the query coverage at 80%. By using the summarize_multiple_blast_outs_into_single_hit_table.pl, lists of sequences passing the filters were created for the 78 resulting permutations.

Building the merged result tables

The lists of all results from each method were merged to avoid redundancy. Since most results were identical, redundancy was removed by adding the sequences in a procedural manner. The order was determined as follows: blastn -> megablast -> tblastx(present for VirDB only) -> bowtie, meaning the base file was in fact blastn, followed by results provided by megablast (that were not in the first file), then results from tblastx that were not found in neither of the two previous files and finally sequences that were only found by bowtie2. The actual merging was achieved with homemade scripts get_single_results_file_for_VirDB_tbx_last.sh, which was employed for results from the VirusDB database and get_single_results_file_per_DB.sh for the remaining cases. This resulted in the creation of a list of contents and the filtered output for each database search using every subset. A subset of the HumanDB aggregate results was created with hits with identity percentage greater of equal to 90%, for human sequence filters.

Using the perl script summarize_multiple_blast_outs_into_single_hit_table.pl, a new table was created with the aggregated tables of all database searches with the human filtered sequences. This table contained a relation of which databases had homologous sequences for each query and how many they were (e.g. a query that only had a hit with the FungiDB would have >0 in the FungiDB column and 0 in the rest). It was parsed using the inhouse script Extract_unique_results.sh to get lists of exclusive hits. By executing it, several mutually exclusive lists were generated, one for human-only, bacterial-only, archaeal-only, PHAST-only, and fungal-only. Viral-only sequences actually allowed for PHAST sequences to be considered as well. Finally, the corresponding search results were extracted using these lists with the script Extract_results_passing_all_filters.sh. After this step, four files contained all results from all database searches: DNA_JnS.out, DNA_split.out, cDNA_JnS.out, and cDNA_split.out.

Contingency table construction

Search results for all queries pointed to representative sequences in clusters, not at the sequences *per se*. However, clustering was reversible as a cluster map was created in earlier steps. The homemade pipeline Unclust_samples_and_process.sh was used to extract each sample's results and get the corresponding taxonomy of each sequence (single or paired) in order to construct a contingency table containing all the results.

The first step of this pipeline was carried out by the inhouse script unclust_sample.pl and consisted in loading the cluster map into memory to extract only those results belonging to a single sample from the aggregated-results files, generating three output files: JnS, PE1 and PE2. These were used, along with the taxonomic information from the databases, to detect which viruses, bacteria, fungi and archaea were present in the samples using another complex homemade script called pick_taxonomy_for_blast_output_2017.pl. This script was written to process samples having either single or paired-end sequences by obtaining a consensus taxonomic assignment using a LCA approach. In all cases, a database-taxonomy and the search output files were required, together with the sample files.

Processing of JnS files was straight-forward. The script loads the taxonomy into memory and read the result for a query sequence, one at a time. It then extracts the corresponding taxonomy for all the matching hits. Whenever one of the hits is clearly much similar to the query than the rest it ignored the others (a difference of 10% identity or more plus equivalent length). Starting at the species level and going up, for each taxonomic rank, a comparison is carried out. If there are two or more different labels, no consensus is reached and the rank is removed. The program then goes up by one taxonomic level and repeats the iteration until the last level has been explored. Ranks showing no consensus are filled with unassigned labels that carry information about the last rank that had a consensus. This was included to avoid unassigned bins to sum taxonomically unrelated items in the contingency tables. In the end, each sequence displays a taxonomy, which may be truncated at the LCA. The output was written as an intermediate file (hits file) containing the counts of all different labels for each taxonomic rank.

The PE1 and PE2 exhibited additional difficulties to deal with. By using the same procedure as for the JnS files, each read for a single sequence is first assigned taxonomy. If both ends produced the same taxonomy, then the script would count one sequence with that taxonomy as it would do for single end reads. If only one had any hits and the other did not, the script wrote the consensus taxonomy of the identified one. If both reads had disparate matches, a new LCA consensus was obtained from the aggregated results and a single count was summed. A series of hits files were produced, containing the aggregated results for each sample in summarized tables.

Finally, using the hit files for all the samples, the script make_OTU_table_from_hit_2017.pl was run to create contingency tables displaying the contents of all samples. The script created seven absolute abundance tables, each one collated at a different taxonomic level, and a corresponding set of relative abundance tables. Two complete sets of tables were generated, one for the DNA set and another for the RNA set. The viral fraction was extracted to new tables and, depending of the type of nucleic acid that was expected, either RNA or DNA viruses were filtered out. PHAST, fungi, bacteria and archaea were also extracted to different tables.

These tables, along with the OTU tables generated for bacteria, comprise the complete set used for diversity analyses.

Standardizing the datasets

The three datasets (cDNA, DNA and 16S) were standardized in a single format, a contingency table containing the total count of each OTU for each sample with each column representing a different sample and each row a taxon. The numerical matrix for all three datasets consists of absolute counts but the WGS (Illumina) sets actually reflect LCAs at the species taxonomic level rather than clusters, contrary to the 16S amplicons which were constructed from closely related sequences. WGS fragments are widespread across the different metagenomes and thus are not expected to cluster but are rather estimates of the presence of certain genome fragments in the sample. On the contrary, in 16S profiles, different OTUs may match the same species and thus appear as can have repeated taxonomies in the same table, something that does not occur with the WGS tables.

The tags in the identified taxonomy bacterial 16S profiles needed standardization, because the classifier truncates taxonomy at different levels. This was achieved by adding empty labels at all missing taxonomic levels to match those in the Illumina sets. In the latter, the taxonomic levels found beyond the LCA are not displayed but rather, a generic undetermined tag "_u" is added at each missing taxonomic level, as well as the last taxonomic level where there was a unique taxonomic node (e.g. the record 1__Bacteria; 2__Proteobacteria; 3__Gammaproteobacteria; 4__Pseudomonadales; 5__Pseudomonadaceae; 6__*Pseudomonas*; 7__*Pseudomonas*_u is missing the species level and has been identified as an undetermined species from the Pseudomonas genus). The undetermined tags are cumulative (e.g. _u_u_u; up to 6 missing ranks can exist) in order to track the last taxonomic level that is available. This was done to avoid merging counts from organisms that had different lineages, a common issue that is often ignored and results in overestimating the number of undetermined items.

It is important to recall that some organisms have not been assigned a scientific name for each node of their taxonomy. Those gaps in the taxonomy were consequently identified and patched up in a similar way to unidentified tags during database construction. As stated before, these nodes had been labelled with "n_ or _n", depending on the closest informative taxonomic rank available. As a result, ranks that are missing due to the current incomplete state of the organism's taxonomy were labelled with "n"s whereas the ranks missing due to a fragment having matching results from phylogenetically distant targets (during homology searches) were labelled with "u"s.

As a result of the methods for database construction, all sequences in the two Illumina datasets had seven taxonomic levels but not all are directly comparable. Living organisms have the following ranks: domain, phylum, class, order, family, genus and species whereas viral results have five taxonomic ranks plus two other non-ranked classifications. These are: virus type, Baltimore classification, order, family, subfamily, genus, species (the first two are non ranked). For the sake of understanding the rest of the process, they will be addressed as taxonomic ranks.

Hereafter, all different taxonomic items in the contingency tables will be addressed as records or OTUs.

Pre-filtering contingency tables

Diversity analyses rely on differences in compositional estimates that in turn are very susceptible to rare record counts. Thus, it is very important to reduce sequencing artefact and items that may have been mislabelled or, in general, records that may not be compared between samples. To achieve this, the contingency tables were trimmed with R to remove records that had less than 10 total appearances (putative artefacts) or were in fewer than two samples. Furthermore, the sequences were ordered by the total sum of counts per record (decreasing order).

The raw Illumina tables (cDNA and DNA sets) contained results from searches against different databases spanning different domains of life including archaea, bacteria, fungi, human, apart from viruses, bacteriophages and prophages. The absolute counts at the domain level were summarized in collated contingency tables for both sets with the QIIME script summarize_taxa.py (--absolute_abundance --level 1). The resulting tables were estimates of total items in the samples from domains Archaea, Eukaryota, Bacteria, as well as prophages, bacteriophages and other viruses (mostly Eukaryotic viruses plus some bacteriophages that are not explicitly classified as such due to missing taxonomic labels). The median values for the domain's totals were obtained with R (control, OL, OSCC, and PVL) and pie charts were created with Microsoft Excel 2016.

Both Illumina sets were then split into different contingency tables, one for each domain or type of virus using the first part of a custom pipeline summarize_tables_only_last_tax.sh. All human hits were removed from the tables. The less abundant sets could not be used separately and thus were added to larger tables. An example of this was the archaea table which was added to the bacteria tables, cDNA_subset_bacteria and DNA_subset_bacteria (the name remained bacteria as the former were not representative). Tables were also created for fungi, prophages and viruses. The latter were assembled differentially in the two datasets. The DNA set did incorporate bacteriophages in the same table. Finally, two collective tables were created including a non-vir table (containing all sequences not in the viral tables, including prophages) and the "all" table. Seven tables were created in total: all, bact, fungi, non-vir, other, proph, and vir. After creating the tables, they were cleansed of columns summing zero using a custom R script, remove_cols_sum_zero.R. Finally, all tables were converted into Biological Observation Matrix (biom) format tables using the biom algorithm (biom convert), a binary format and QIIME's default.

Compositional Analysis

The composition of the tables was assessed by executing a homemade pipeline, summarize_tables_only_last_tax.sh, which uses QIIME's summarize_taxa.py script and inhouse R scripts. Then, for each table, it aggregates the absolute counts for each of the seven taxonomic levels (--level 1,2,3,4,5,6,7 and –absolute_abundance), effectively creating new tables with recalculated total counts. Each of these is then sorted by total sum using an R script, sort_summarized_table_dont_reformat.R and the taxonomic label is trimmed to only bear the last level in each table (species in table 7, genus in table 6, etc.). Finally, the homemade script graph_composition.R creates a stacked-bars plot (data represented as percentages) for each taxonomic level, sorted by decreasing abundance average and bearing additional information about the sample depth (differences in total sequences per subject). The same process was carried out with the 16S profile data.

Alpha diversity and sample rarefaction

Assessing the within-sample variability (alpha diversity) allows estimating the sampling effort, the richness and dispersion of the data. To achieve this, different diversity indexes can be calculated from the contingency tables. However, in order to compare these values, the sampling unevenness must be tacked by rarefying (resampling).

To compare alpha diversity across a whole set of samples, an equal sample depth (total counts) is required. This is achieved by resampling the contingency tables and can be repeated to provide statistical support. This was thus not feasible with most subset tables. Therefore, for the Illumina datasets, the "subset_all" tables were used for calculations. Resampling was carried out using the alpha_rarefaction.py workflow from QIIME. The size of the fifth largest sample to avoid plotting to the scale of the largest sample, which is usually much larger than the rest. This value, the maximum value (-e), and the minimum was set to 100 (--min_rare_depth 100). The steps were set to 50 (-n 50), each with an increasingly larger sample size that is calculated automatically. The repetitions were set to 20 (multiple_rarefactions:num_reps 20 via parameters file). The metric was set to observed_species (alpha_diversity:metrics observed_otus).

The pipeline draws multiple rarefactions with increasing larger total items to evaluate how records accumulate changing progressively (separately for each sample). Since the resampling is carried out with no replacement, samples are dropped from the evaluation once the total items value becomes larger than their maximum number of items. Each step has multiple rarefactions occurring as sampling variation must be considered. For each rarefaction, the estimator (observed counts) is evaluated and the average is obtained in collated tables. These data were used to plot a set of rarefaction curves with the mean values of the observed items per sample using a custom R script Rarefaction_curves.R.

Additionally, in order to assess the differences in alpha diversity of each sample, a richness estimator and a diversity index were used for the Illumina subsets.

The Chao1 estimator (*330*) was calculated to evaluate richness per sample. This estimator depends on the number of low abundance observations.

$$S_1 = S_{obs} + \frac{F_1^2}{2F_2}$$

In the formula, *Sobs* is the number of species in the sample, *F1* is the number of singletons (species with only a single occurrence in the sample) and *F2* is the number of doubletons (the number of species with exactly two occurrences in the sample). It is expected that rare species are recovered only once until sampling increases up to the point of saturation, where most species have been sampled at least twice, meaning only few observations are expected beyond this point.

The Shannon's diversity index (*331*) was also used to evaluate entropy (evenness) and abundance of the records in the tables.

$$H = -\sum_{i=1}^{s} (p_i)(\ln p_i)$$

In the formula, the proportion of species *i* is calculated relative to the total number of species p_i, then multiplied by the natural logarithm of this proportion (ln p_i). The resulting product is summed across species, and multiplied by -1.

The idea is that prediction of the next item is easy in skewed populations with few different items. The more different and evenly distributed these are, the less probable it is to predict the next item. The effective number of species can be estimated as exp(H), which reflects an estimate of the real biodiversity.

Using QIIME's multiple_rarefactions.py script, 1,000 rarefactions were carried out at 3,000 (-x --num_reps 1000 -m 3000 -x 3000; min and max are the same). The alpha diversity was calculated with the script alpha_diversity.py with the Chao1, Shannon and observed otus (--metrics chao1, shannon, observed_otus). The results of each rarefaction's estimates were aggregated using QIIME's collate_alpha.py script. Boxplots for each samples rarefied Chao1 and Shannon estimates were created using a homemade R script, boxplots.R. The script compare_alpha_diversity.py was then used to compare the richness and diversity by groups of samples (control, OL, OSCC, and PVL) and estimate if they were significantly different using a non-parametric two-tailed t-test using 1,000 Monte Carlo simulations with Benjamini-Hochberg's false discovery rate (FDR)correction (-t nonparametric, -n 1000, -p fdr) (*332*).

Beta diversity

Alpha rarefaction evaluates the diversity within samples, whereas beta diversity assays the variation between different samples and groups. This is normally achieved by calculating a distance matrix reflecting the differences between each pair of samples using a diversity index. Linear algebraic combinations are then calculated in a multidimensional scaling process such as Principal Coordinates Analysis (PCoA) for explaining variation in the matrix. This consists on modelling variation as a combination of contributions of the OTUs to reduce the complexity of observations. The resulting linear combinations (eigenvalues) represent the aggregated factors that contribute towards dissimilarities and can be visualized as a coordinate map of all samples and may be driven by the composition, the group they belong to, etc.

Diversity analyses are affected by the sparsity in the contingency table (the proportion of zero values). Rare items occurring in only a few samples influence differences in dissimilarity matrix calculations. Furthermore, in abundance-based analyses, for samples to be comparable, items with differential abundance but present in several samples are more useful. For these analyses, the 16S summarized table at the seventh level was used (collated at the species level). From the cDNA and DNA sets, the subsets containing viruses and the ones containing non-viruses (the rest) were selected. Since most of the contents are bacteria, for convenience, these sets will be referred to as bacterial. In order to reduce data sparsity, low-occurrence items were removed from the contingency tables using R. Only records occurring in at least 10% of the samples were kept. Additionally, samples containing totals that were much smaller than the rest were removed (determined empirically for each table as the minimum varied). Hereafter, the resulting sets will be addressed as strict sets (e.g. DNA_strict).

Due to the large variation in the total number of items in each sample, an adequate comparison depends on standardization of the total counts. This can either be achieved by transforming the data to statistically similar distributions (adjust variance, mean, quantiles, apply log or factor, etc.) of by rarefying the samples so that counts are even. One of the most common transformations is to scale the content in terms of the total sums (converting into relative numbers). However, this leads to the magnification of differences that are due to sampling issues and differential sampling efforts. Instead, newer normalization techniques take into account inner variation in the metagenome for working with data with high sparsity. One of these techniques is cumulative sum scaling (CSS) from the metagenomeSeq package in R (*333*). In this normalization technique, counts are divided by the cumulative sum of counts up to a percentile that is determined using a data-driven approach. Non-viral datasets were CSS-normalized using QIIME's normalize_table.py (-a CSS) which also applies a logarithmic transform. However, for the low-abundance viral datasets, rarefaction was used instead as variation due to sampling effort was higher. Rarefactions were carried out with the script single_rarefaction.py with variable depth. The beta diversity was calculated with parts of the QIIME's pipeline beta_diversity_through_plots.py. Specifically, rarefactions were avoided for normalized datasets because no double transformation is required for the data.

The Bray-Curtis (*334*) dissimilarity index was calculated.

$$dBC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

In the formula, dissimilarity between objects i and j is expressed as one minus the ratio of common items in the two sets, C_{ij} and specific specimens, S_i and S_j. It is used to quantify the compositional dissimilarity between two different samples. The measure ranges from 0 to 1, with 0 meaning identical and 1 meaning completely separated populations. Although not a distance, dissimilarity can be used to evaluate differences between samples given their abundance and composition.

The Jaccard binary index (*335*) was also calculated.

$$dJ(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

In the formula, the Jaccard distance is simply the ratio of the unique items in the distributions and the total number of items in both. It measures dissimilarity between finite sample sets evaluating presence or absence instead of abundance. It evaluates n binary attributes between two samples.

Statistical significance of sample groupings was calculated using the resulting distance matrices with the script compare_categories.py using Adonis (*336*), a nonparametric analysis of variance. This splits a distance matrix among sources of variation in order to describe the strength and significance that a variable has over distance variation. Groups (control, OL, OSCC, or PVL), age, sample site location and gender were used to detect statistical relevance in both the Jaccard and Bray-Curtis distance matrices of all beta diversity analyses of the tables. Additionally, boxplots were created using the script make_distance_boxplots.py to compare distances between categories. Two-sample t-tests (parametric) and Monte Carlo (nonparametric) tests were carried out to assess the significance of differential distributions, including Bonferroni corrections. PCoA were executed with principal_coordinates.py and plotted with Emperor (*337*) visualization tool using the script make_emperor.py with groupings by categories in the map file. The PCoA axes were plotted against the depth as well to verify the effect of sampling effort on the distributions (--add_vector Depth).

Differential Analysis

In order to evaluate which OTUs were causing separations in the PCoAs, several methods were employed. Using the same tables that were employed during the construction of the distance matrices (the strict tables, prior to normalization/rarefaction), heatmaps were produced to visualize data abundance. This was achieved with the homemade pipeline Heatmaps.sh, which uses the summarize_taxa.py script to obtain the seven level tables for each taxonomic rank, then uses the inhouse script sort_summarized_table_dont_reformat.R to filter all ranks, except the highest in each case, and finally constructs the heatmap with make_otu_heatmap.py using grouping by group (control, OL, OSCC, or PVL).

Statistical differences were evaluated both in groups and in paired comparisons. For the former, the QIIME script group_significance.py was used to compare record frequencies across sample groups to find which of the items had the highest probability of being differently represented depending on the sample Group, Age, Gender and Sampling location categories (-c) using Kruskal-Wallis (non-parametric; -s kruskal_wallis). Corrections were tested with Bonferroni and FDR for multiple comparisons. Additionally, negative binomial Wald test was used to identify which records were differentially abundant across two sample categories (paired comparisons). Originally developed for RNA-seq, the method assumes a Gamma-Poisson distribution to estimate the probability of extreme events (large fold changes that just appear by chance) for discrete replicates. This was achieved with the strict datasets as transformations mask the potentially interesting markers. using the differential_abundance.py script to run DESeq2 (*338*) with all permutations of the Group (Control, OL, OSCC, and PVL) category set (-a DESeq2_nbinom -c Group).

Finally, the online-based tool Linear Discriminant Analysis Effect Size (LEfSe) (*339*) was used to find relevant differentially distributed markers in the groups. The algorithm depends on the hierarchical structure of the contingency table with different taxonomic levels to statistically compare (using Kruskall-Wallis tests) features that are differentially distributed between the different clusters of samples (e.g. groups in the study). The subsets are then used to build vectors with linear combinations similar in concept to PCoA's axes but trying to model the differences between the classes. The vectors are then used to rank the features and calculate the effect size over the classes (the contribution to their difference). This was executed with the strict (pre-normalization/rarefaction) using default parameters using the strict datasets (pre-normalization/rarefaction data).

RESULTS

Screening for oncoviruses

Samples

A total of 40 samples of variable sizes and tissue characteristics were provided for the retrospective study (Table 11). Size and tissue composition of the biopsies varied greatly between samples. OSCC samples were in general larger and softer than controls whereas the hyperkeratotic patches from OL and some PVL samples were the most difficult to disrupt.

Table 11. Samples from the oncoviruses screening retrospective study.

Sample	Group	Age	Sex	Location (additional locations)	Clinical appearance
C01	CTRL	32	F	Third molar extraction area	Normal
C02	CTRL	29	F	Third molar extraction area	Normal
C03	CTRL	18	H	Third molar extraction area	Normal
C04	CTRL	24	F	Third molar extraction area	Normal
C05	CTRL	26	F	Third molar extraction area	Normal
C06	CTRL	25	H	Third molar extraction area	Normal
C07	CTRL	21	F	Third molar extraction area	Normal
C08	CTRL	26	F	Third molar extraction area	Normal
C09	CTRL	22	F	Third molar extraction area	Normal
C10	CTRL	26	H	Third molar extraction area	Normal
L01	OL	55	F	Tongue	White patch
L02	OL	59	F	Cheek	White patch
L03	OL	72	F	Tongue	White patch
L04	OL	54	F	Mouth floor	White patch
L05	OL	65	F	Cheek	White patch
L06	OL	71	F	Tongue	Ulcer
L07	OL	68	H	Tongue	White patch
L08	OL	54	F	Mouth floor	White patch
L09	OL	83	F	Tongue	White patch
L10	OL	74	F	Gingiva	White patch
S01	OSCC	84	H	Tongue	Exophytic tumour
S02	OSCC	80	H	Mouth floor	Exophytic tumour
S03	OSCC	84	H	Tongue	Ulcer
S04	OSCC	82	F	Gingiva	Ulcer
S05	OSCC	53	H	Tongue	Exophytic tumour
S06	OSCC	52	F	Mouth floor	Erythroplasia
S07	OSCC	92	F	Gingiva (palate)	Exophytic tumour
S08	OSCC	62	H	Gingiva	Ulcer
S09	OSCC	68	H	Tongue (mouth floor)	Ulcer
S10	OSCC	64	H	Tongue	Ulcer
V01	PVL	64	F	Gingiva	Verrucous patch
V02	PVL	76	F	Cheek (tongue, mouth floor, gingiva)	Verrucous patch

V03	PVL	83	F	Gingiva (tongue, mouth floor, cheek)	Verrucous patches
V04	PVL	75	F	Gingiva	White patch
V05	PVL	65	F	Mouth floor (tongue, gingiva)	Tumour
V06	PVL	80	F	Tongue (mouth floor, gingiva y palate)	White patch and erythroplasia
V07	PVL	67	F	Gingiva	Verrucous patches
V08	PVL	57	F	Cheek (tongue, mouth floor, gingiva, palate	Verrucous patches and erythroplasia
V09	PVL	71	F	Gingiva (cheek, tongue)	Verrucous patches
V10	PVL	65	F	Gingiva (tongue, mouth floor, palate, cheek)	Verrucous patches

All samples were processed under the following the same protocol. After mechanical tissue disruption, container tubes presented a light red-brown colour and small fragments of tissue that managed to avoid complete homogenization. These were discarded.

DNA extraction was successful for 38 samples in total. Twenty-five of them were correctly amplified during the first 16S rDNA amplification round. DNA from the 13 remaining samples was effectively enriched with GenomiPhi, as confirmed by the subsequent 16S rDNA amplification. Two samples were ultimately discarded from subsequent analyses as no genomic DNA was confirmed to be present in any of them. Until this point, the study was carried out as a blind exploration. After their origin was disclosed, they were separated in groups of oral leukoplakia (OL), oral squamous cell carcinoma (OSCC), proliferative verrucous leukoplakia (PVL), and a healthy group (CTRL). The two samples that failed belonged to the OSCC and the healthy control group.

Screening

The papillomavirus screening with primers FAP59 and FAP64 resulted in the amplification of two fragments, in 32 of samples, all confirmed by gel electrophoresis (Figure 44). The fragments were between 300 and 400 bp in length, contrasting to the expected 600 bp of the theoretical papillomavirus amplicon. Furthermore, none of them resembled the band in the positive control. A band from each size was taken as representative and was loaded in a new gel for excision, purification and sequencing.

The second primer set for papillomaviruses, consisting of CP4 and CP5 primers, generated a single band in 33 amplifications as seen in gel electrophoresis (Figure 45). This fragment was approximately 250 bp long, contrary to the expected ~600 bp band in the positive control. A sequence of similar size was detected in a sample from oral leukoplakia. The megablast alignment against the family of papillomaviruses yielded no results for any of the sequences detected in the bands, except for the positive control, which matched *Eidolon helvum papillomavirus* E1 gene (evalue 0.0 and 100% identity), as expected. No significant hits were found with tblastx, against the papillomavirus database. For the band that was present in most samples, blastn alignments against a sequence of *Homo sapiens* steroid receptor RNA activator 1 were reported (86% query coverage, evalue 5e-28, and 80% identity). Blastn also produced a short local alignment with *Leuconostoc mesenteroides* for the ~600 bp band with relatively low evalue (1e-3 and 85% identity over a span of <20% of the total read length).

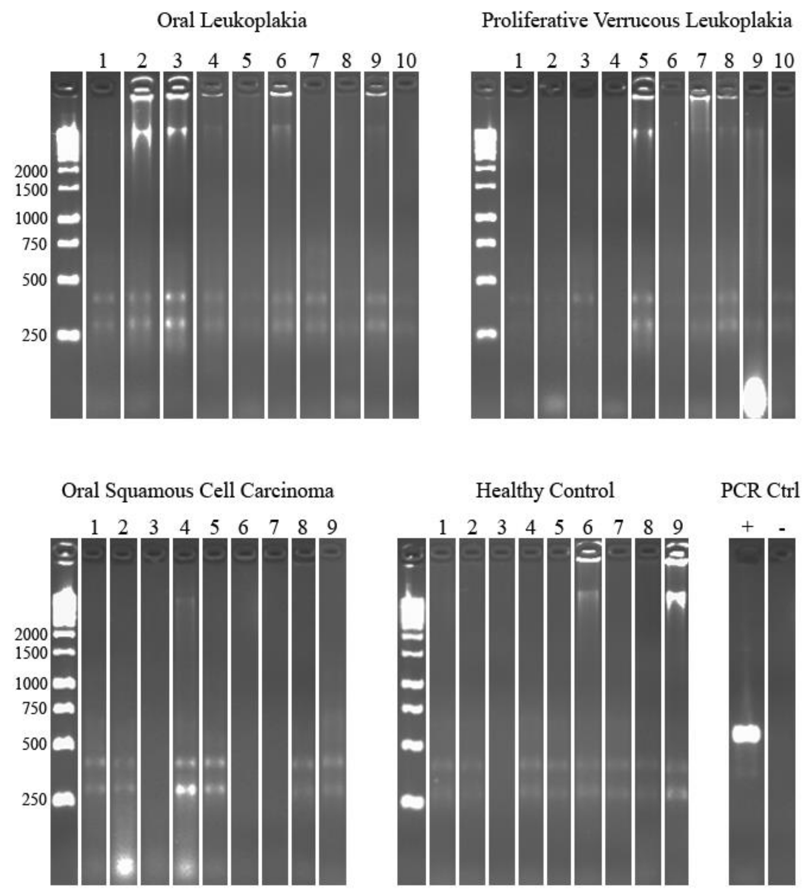


Figure 44. Papillomavirus screening with primers FAP59 and FAP64. Samples containing DNA extracted from biopsy samples belonging to OL, PVL, OSCC and control groups were stained with GelRed, loaded in 1.4% agarose gel for electrophoresis and photographed under a UV light. Exposure was increased to capture a sharper image of the bands. DNA from *Eidolon helvum papillomavirus* was used as a positive control. Adapted from García-López *et al*., (*210*).

Polyomavirus screening with primers VP1-2f and VP1-2r resulted in the detection of multiple bands in the gel with highly variable lengths (Figure 46). The most prominent band was from fragments that migrated below the 500 bp marker. Also, a fragment in the OSCC group matched was detected in the 250-270 bp range, the same as the expected amplicon size. Representative samples for each band size were purified and sequenced. None of the resulting sequences yielded any hits with neither megablast nor tblastx against the polyomavirus database. Blastn produced partial alignments for the ~250 bp band with *Homo sapiens* chromosome 5 clone CTB-99P17(evalue 9e-55 and 99% identity) and for the ~800 bp fragment, matching an undetermined sequence from *Homo sapiens* chromosome 19 (evalue 0.0 and 95% identity). The ~700 bp fragments aligned with a sequence from *Homo sapiens* protein kinase C, epsilon gene (e-value 0.0 and >87% identity). Also, the ~450 bp fragment, seen in gel electrophoresis as a faint band, was aligned to a sequence on chromosome 12 (evalue 7e-145 and 95% identity). Finally, the ~300 bp fragment sequence was aligned to sequence *Homo sapiens* chromosome 5 clone CTB-55A14 (evalue 6e-13 and 85% identity).

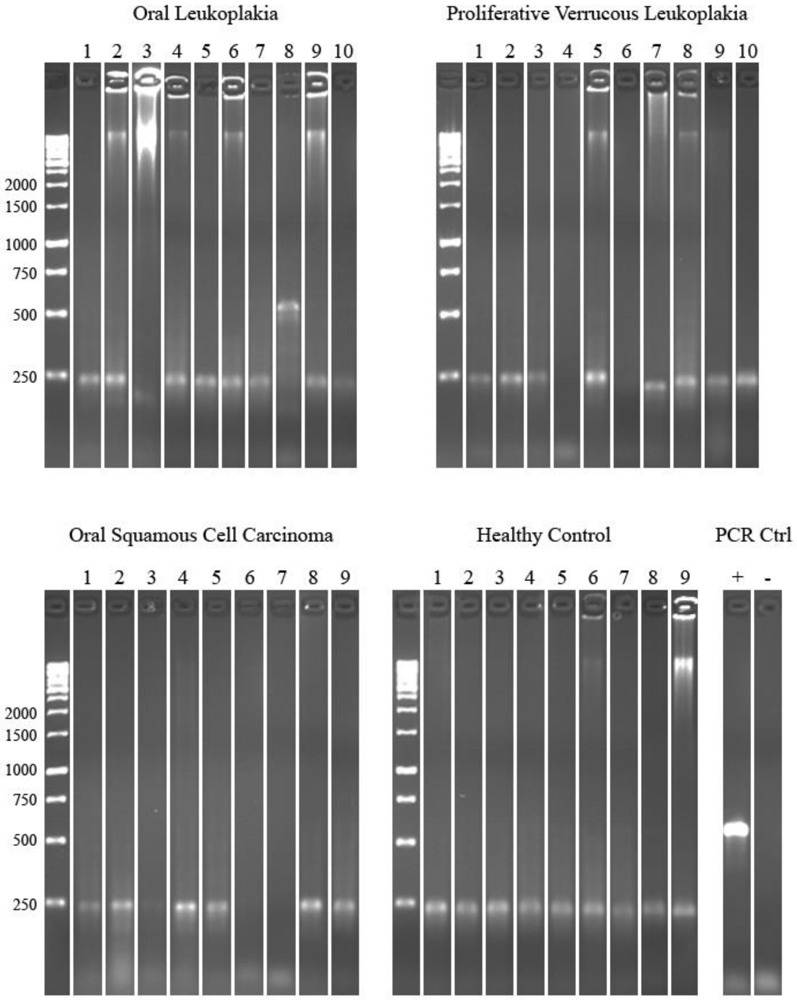


Figure 45. Amplicon yield of PCR with primers CP4 and CP5 for papillomavirus amplification. Adapted from García-López *et al*., (*210*).

The amplifications using primers MCVPS1f and MCVPS1r for Merkel Cell polyomaviruses resulted in a band of ~320 bp in most samples as seen in the gel electrophoresis, as well as several other bands that correspond >700 bp fragments (Figure 47). The expected band for the control amplicon was expected to be ~110 bp long. Thus, bands larger than 800 bp were not considered for sequencing. Representative bands with sizes ~320 and 700 bp were subjected to sequencing. All reads were aligned with megablast and tBLASTx against the *Polyomaviridae* family database but yielded no significant results. Homology searches with BLASTn against the nr database resulted in an alignment of the 320 bp fragment with *Homo sapiens* PAC clone RP4-539M6 (evalue 1e-104 and 98% identity) whilst the sequence from the larger 700 bp fragment aligned a sequence in *Homo sapiens* chromosome 1 (evalue 0.0 and identity 99%).

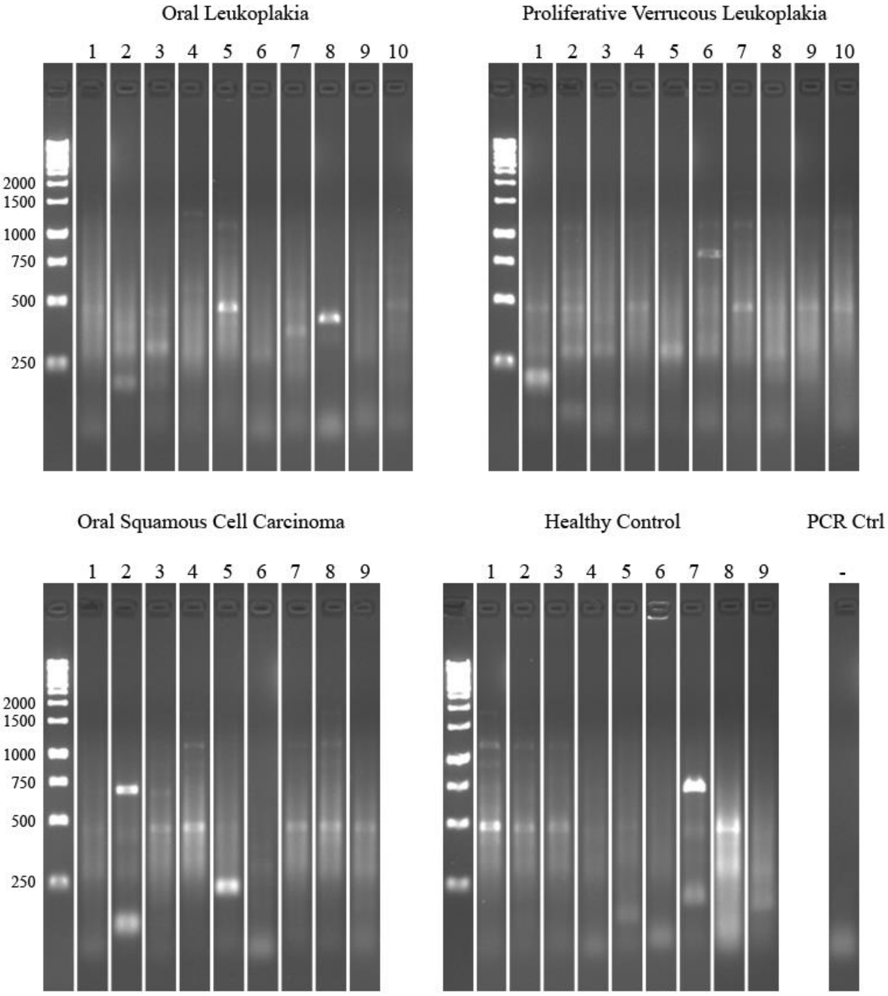


Figure 46. Polyomavirus screening. Amplicon yield of the second round of the nested PCR with primers VP1-2f and VP1-2r for universal polyomavirus amplification. Exposure was increased to capture a sharper image of the bands. Adapted from García-López *et al*., (*210*).

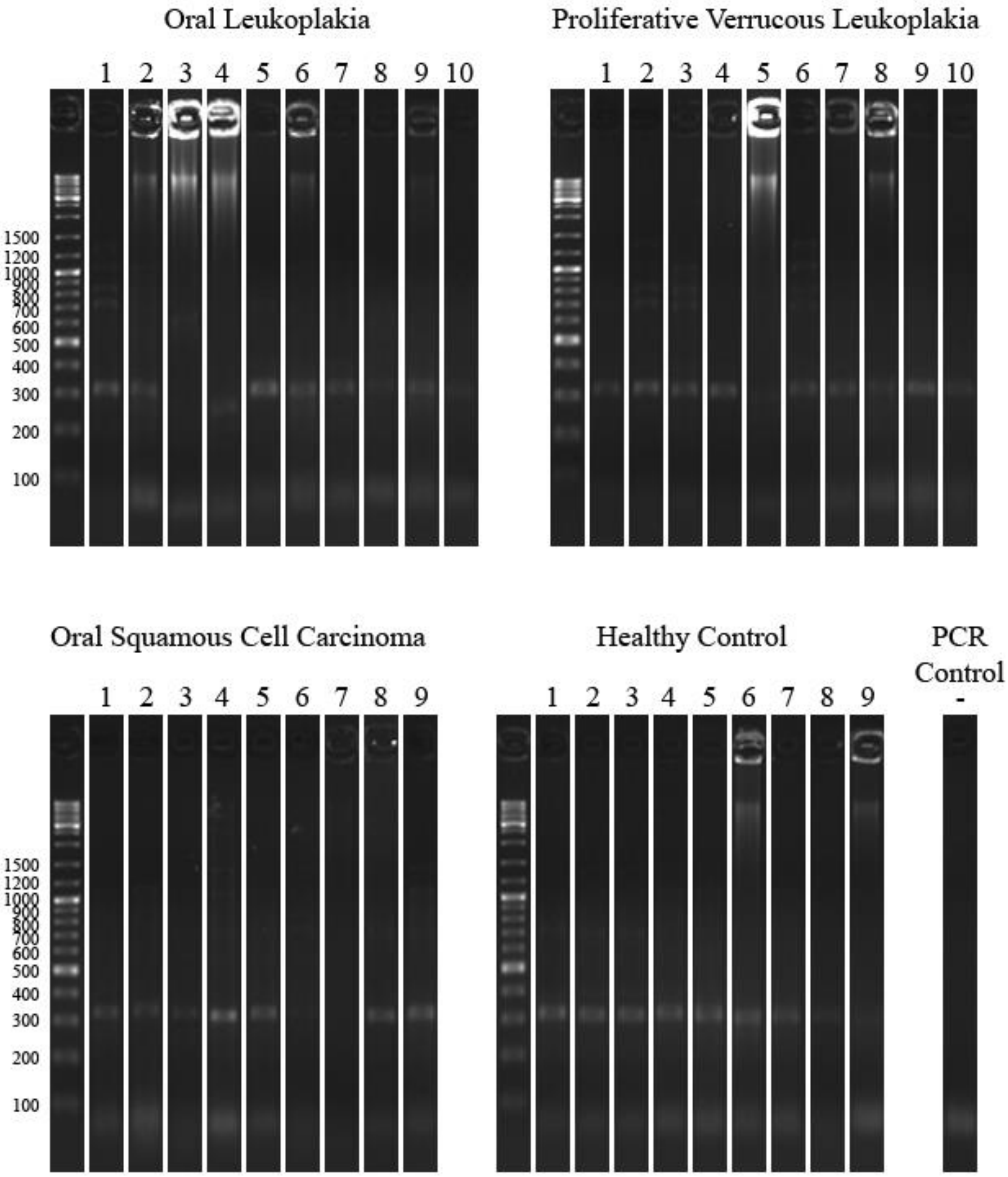


Figure 47. Amplicon yield of PCR with primers MCVPS1f and MCVPS1r for Merkel cell polyomavirus amplification. Adapted from García-Lopez *et al.*, (*210*).

The last set of primers (DFAf, ILKf, and KG1r) for screening herpesvirus resulted in the amplification of one ~300 bp faint band in most samples when run in gel electrophoresis. Contrastingly, the positive control yielded a 250 bp band that was aligned with megablast search against the *Herpesviridae* family database with *Human herpesvirus* 8 isolate KSHV (evalue 2e-69 and 97% identity). The other bands yielded no significant results for either type of blast.

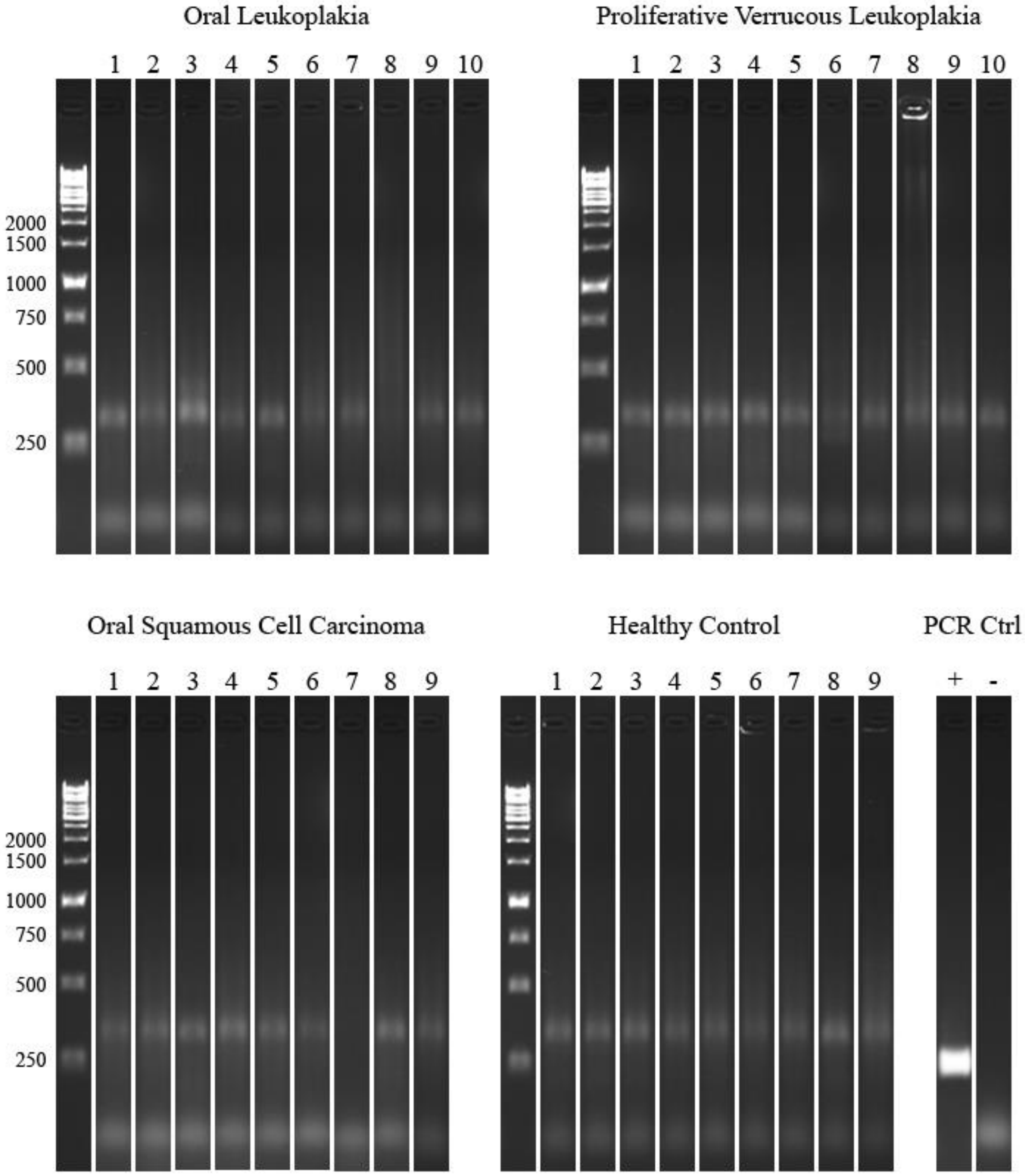


Figure 48. DNA from Epstein–Barr virus and Kaposi's sarcoma-associated herpesvirus was used as positive control. Adapted from García-López *et al*., (*210*).

Metagenomic and 16S profiling study

Samples

A total of 41 samples were used for this second retrospective study. Sizes and tissue hardness varied among the different biopsies (Table 12).

Table 12. Samples from the metagenomic and 16S profiling retrospective study.

Sample	Group	Age	Sex	Location (additional locations)	Clinical appearance
C01	CTRL	32	F	Third molar extraction area	Normal
C02	CTRL	23	F	Third molar extraction area	Normal
C03	CTRL	20	F	Third molar extraction area	Normal
C04	CTRL	23	M	Third molar extraction area	Normal
C05	CTRL	19	M	Third molar extraction area	Normal
L01	OL	37	M	Gingiva	Homogeneous Leukoplakia
L02	OL	46	F	Upper gingiva (lower gingiva)	Homogeneous Leukoplakia
L03	OL	59	F	Buccal mucosa	Homogeneous Leukoplakia
L04	OL	64	F	Gingiva	Homogeneous Leukoplakia
L05	OL	45	F	Palatine gingiva	Homogeneous Leukoplakia
L06	OL	48	M	Tongue-ventral surface	Homogeneous Leukoplakia
L07	OL	59	M	Mouth floor	Homogeneous Leukoplakia
L08	OL	82	F	Upper gingiva	Homogeneous Leukoplakia
L09	OL	89	F	Tongue-lateral border (mouth floor, buccal mucosa)	Verrucous Leukoplakia
L10	OL	74	M	Lower lip	Homogeneous Leukoplakia
S01	OSCC	44	M	Tongue-lateral border	Ulcer
S02	OSCC	82	M	Retromolar trigone	Ulcer
S03	OSCC	86	F	Upper gingiva	Ulcer
S04	OSCC	82	F	Lower gingiva	Ulcer
S05	OSCC	71	M	Tongue-ventral surface (mouth floor)	Ulcer
S06	OSCC	53	M	Mouth floor	Ulcer
S07	OSCC	95	F	Upper gingiva	Tumour
S08	OSCC	87	F	Tongue (lateral border)	Ulcer
S09	OSCC	64	M	Lower lip	Ulcer
S10	OSCC	76	M	Retro-commissural mucosa	Erythroplasia
S11	CTRL	65	M	Lower gingiva (tongue-lateral border,mouth floor, retromolar trigone)	Erythroplasia, ulcer, homogenous and non-homogenousleukoplakia
S12	CTRL	81	F	Lower gingiva (upper gingiva, palate)	Erythroplasia, exophytic lesion, non-homogenous leukoplakia
S13	CTRL	84	F	Tongue (lateral border)	Ulcer
S14	CTRL	85	M	Lower gingiva	Ulcer
S15	CTRL	93	F	Palate (lower gingiva)	Ulcer
S16	CTRL	89	F	Tongue-lateral border	Erythroplasia
V01	PVL	60	F	Tongue (gingiva)	Verrucous Leukoplakia
V02	PVL	84	F	Upper gingiva (buccal mucosa, tongue, mouth floor)	Verrucous Leukoplakia
V03	PVL	80	F	Lower gingiva (buccal mucosa, palate, lower lip)	Verrucous Leukoplakia
V04	PVL	79	F	Upper gingiva (lower gingiva, mouth floor, tongue - lateral border, buccal mucosa)	Verrucous Leukoplakia
V05	PVL	50	F	Gingiva (vestibular gingiva 4 quadrants)	Leukoplasia - No dysplasia
V06	PVL	60	M	Lower gingiva (mouth floor, buccal mucosa)	Verrucous leukoplakia+erythroplasia
V07	PVL	79	M	Lower gingiva (upper gingiva, buccal mucosa)	Verrucous Leukoplakia
V08	PVL	76	M	Upper gingiva (lower gingiva, tongue)	Verrucous Leukoplakia
V09	PVL	59	F	Lower gingiva (upper gingiva, palate, buccal mucosa, mouth floor, tongue)	Verrucous leukoplakia+erythroplasia
V10	PVL	77	F	Lower gingiva (palate, buccal mucosa, mouth floor)	Homogeneous Leukoplakia

Tissue disruption was carried out for all samples. Most hard biopsies (i.e. keratotic patches were considerably more difficult to homogenise and required more time in the bead beating disruption step. Container tubes presented a light red-brown colour afterwards and tissue pieces always remained after the process.

Filtering and DNA extraction with commercial kits was successful for all 41 samples, as confirmed by the successful amplification of a fragment matching the size (~550-650 bp) of the V1-3 region of the16S rDNA with primers E8F and B530R (Figure 49).

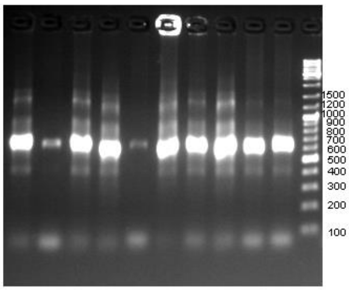


Figure 49. Example of the 16S amplification confirmed in agarose gels. An intense band slightly bigger than 550 bp is detected in all correct amplification results due to the addition of the MIDs (barcodes).

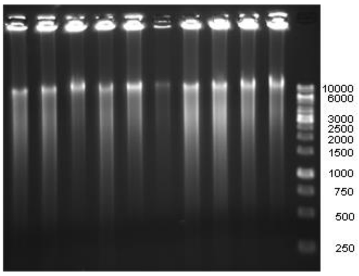


Figure 50. Example results of GenomiPhi enrichment. Genomic DNA is detected as a large smear spanning fragments of different lengths in a continuous distribution that concentrates on larger amplifications. Contrary to PCR, amplification is not specific in MDA.

The viral extraction was amplified with MDA, resulting in the enrichment of total gDNA consisting in large fragments (over 10,000 bp). This was confirmed with gel electrophoresis (Figure 50). All samples were successfully amplified.

Retrotranscription of the RNA fraction was carried out with the SISPA protocol and the resulting cDNA was confirmed via gel electrophoresis after amplification using the SISPA primers, seen as a smear concentrated on shorter fragments. Sample V03 (PVL group) was never amplified and was ultimately removed from the study. All the remaining samples were confirmed and loaded into gels for selecting the fragments of approximately 250-500 bp via band excision and subsequent purification, which was later confirmed with new gel electrophoreses (Figure 51).

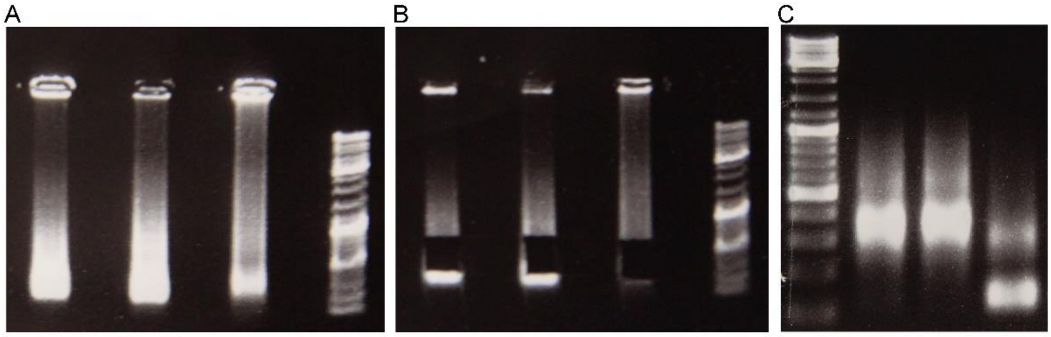


Figure 51. Example of SISPA amplification results. A) A smear rich in short sequences is shown in SISPA samples in gel electrophoresis. B) The fragments spanning 250-500 bp were removed with sterile scalpels. C) A new gel is run. Some larger or shorter fragments are still present in the new gel but it is mostly the fraction that can be sequenced.

Sequencing

For the 16S set the 454 pyrosequencing produced a total of 264 Mbp in four eights of a plate (summing 62, 63, 68, and 70 Mbp respectively). A total of 588,225 sequences with a raw mean of 448 bp. For the DNA, the Illumina MiSeq generated a total of 9.40 Gbp with 16,400,917 2x300 paired-ends sequences in total. For the cDNA dataset, the Illumina MiSeq generated a total of 2.04 Gbp of 2x200 paired-end sequences with 10,181,292 pairs in total.

The total numbers of identified sequences are shown in Figure 52, showing the total numbers of identified sample in the 16S, the DNA and the cDNA set.

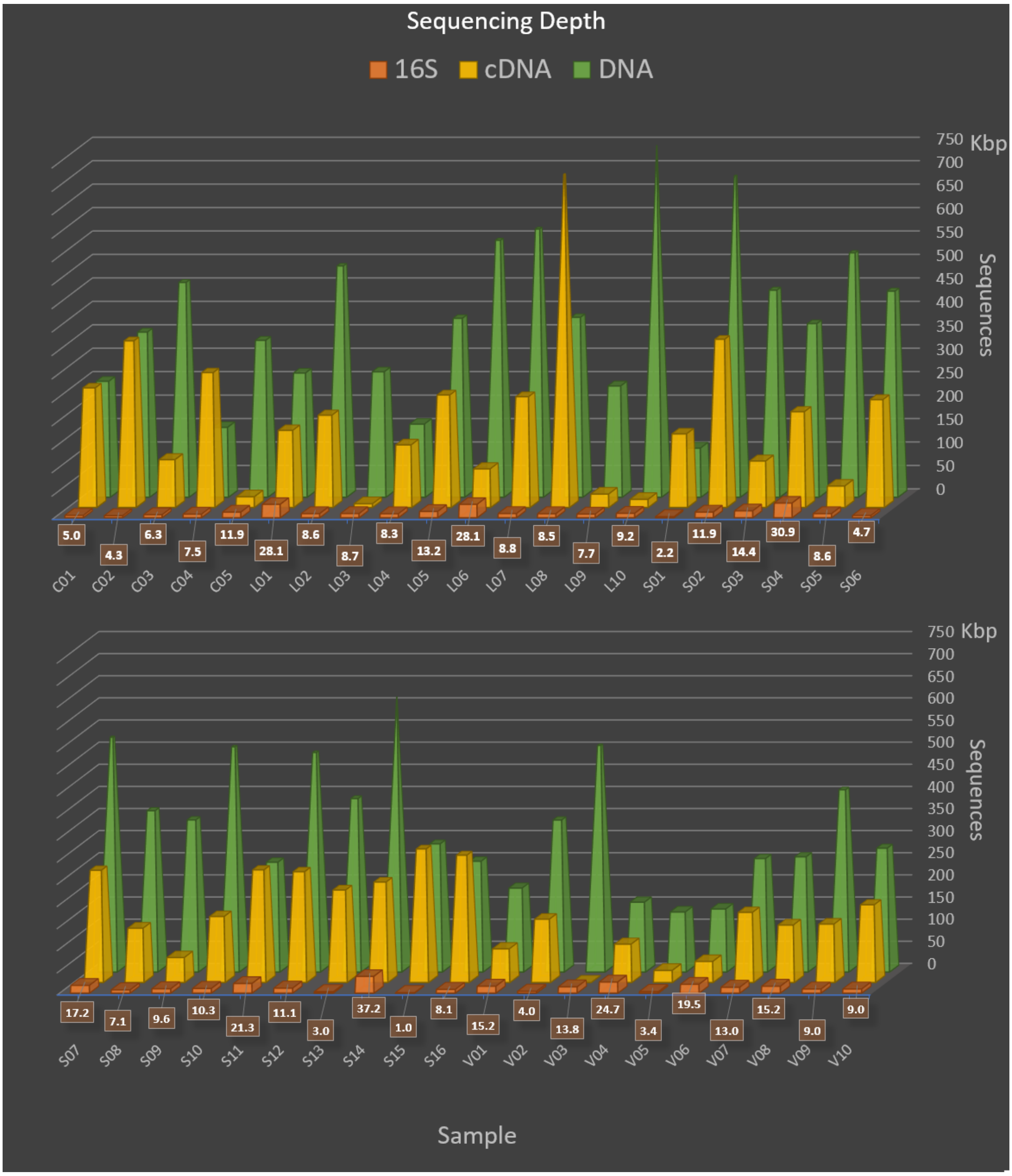


Figure 52. Sequencing depth per sample. The total maximum counts (in Kbp) for each sample is shown for the three datasets. For the Illumina datasets, the maximum is achieved after the sequence scavenging pipeline, by recovering reads from the undetermined pool. In the 16S set, the sequences are split after quality filters and thus vary from the potential maximum. Detailed totals are provided for the 454 set as it has the lowest values of the three.

The 16S had the highest quality of the three sets and it was processed as a single dataset. Quality trimming and filters produced a total of 528,127 reads that were split into the different samples as shown in Table 13. In total, 7.82 % of the sequences remained unidentified as they did not cluster with the reference OTUs bearing a valid taxonomy with at least 97% identity.

Table 13. Summary of the 16S set processing.

Raw		Qual. Filters		Identified		Unidentified	
588225		528127		489814		38313	
CTRL		OL		OSCC		PVL	
C01	5039	L01	28121	S01	2221	V01	15246
C02	4289	L02	8553	S02	11891	V02	3961
C03	6344	L03	8744	S03	14413	V03	13802
C04	7536	L04	8311	S04	30923	V04	24729
C05	11885	L05	13172	S05	8599	V05	3397
		L06	28109	S06	4677	V06	19542
		L07	8812	S07	17155	V07	13005
		L08	8464	S08	7133	V08	15211
		L09	7718	S09	9632	V09	9031
		L10	9203	S10	10326	V10	8998
				S11	21280		
				S12	11058		
				S13	2995		
				S14	37157		
				S15	998		
				S16	8134		

These sequences clustered at 97% identity with the usearch_qf pipeline using a 16S rDNA reference, forming a total of 4,302 different OTUs with at least two samples that were assigned a taxonomy based on their homology to existing clusters of bacteria. The complete resulting contingency table can be found in the online repository https://github.com/rodrigogarlop/PVL inside the raw tables folder.

The results of the DNA processing are presented in Table 14, which summarizes the total sequences in each step. These include dereplication, scavenger (recovery of sequences from the unidentified pool), adaptor trimming, filtering and a new dereplication and the joining of the PE. The resulting sets, JnS and Split, are in the last columns with the split referring to total pairs. Sequences in the Unidentified pool (U) are only shown for the first columns, after which they become irrelevant. The most critical steps were the quality filters, which removed a large proportion of samples.

Table 14. Summary of the DNA set processing.

Sample	Original	Dereplicate	Scavenger	Adaptors	Trim, Filter, dereplication	Joined	JnS	Split
C01	244963	244923	246340	244988	216186	188306	193183	27880
C02	349016	349013	350887	344639	278027	251173	259007	26854
C03	454020	453822	455856	450843	353705	306702	314110	47003
C04	147664	147647	148463	147854	130394	109667	111969	20727
C05	331569	331559	333135	328817	275263	245734	252697	29529
L01	262576	262576	263383	259119	206701	163165	183562	43536
L02	488812	488789	492037	471375	360920	302703	321098	58217
L03	265116	265114	266418	262170	221892	179244	190711	42648
L04	154471	154471	154671	152921	104963	73085	107341	31878
L05	378009	378008	379190	375368	310740	238906	264918	71834
L06	543089	543085	547312	522664	383259	339455	360124	43804
L07	566779	566732	570468	536839	362065	305996	324846	56069
L08	380102	380102	381843	367463	265427	225040	253001	40387
L09	235254	235254	236882	223843	141770	121808	130918	19962
L10	744260	744259	750020	702085	475441	432309	455369	43132
S01	103756	103753	104499	102626	84700	74657	77606	10043
S02	679025	679020	684358	659394	497606	451429	468662	46177
S03	437633	437622	440906	421780	259467	236854	247034	22613
S04	366815	366703	369002	358788	273387	239680	246849	33707
S05	516323	516315	519885	508166	393279	354714	370139	38565
S06	435531	435524	438137	432728	359898	306045	322949	53853
S07	524107	524099	527580	520877	420948	361382	380170	59566
S08	360071	360067	363438	354513	268397	240628	251878	27769
S09	339588	339588	341104	337325	105918	90923	105685	14995
S10	502678	502667	505822	500615	285445	225669	249390	59776
S11	245111	245105	246652	242072	188321	166145	170178	22176
S12	489970	489964	493714	466467	329986	296329	309558	33657
S13	387217	387216	389020	378742	295892	263863	273763	32029
S14	615071	609853	613405	602916	371036	319684	325274	51352
S15	286054	286054	287629	279873	221893	196042	202605	25851
S16	247352	247350	248641	240870	185014	164592	171280	20422
V01	187492	187491	188419	187213	171290	143744	149778	27546
V02	339247	339221	340528	338452	295891	255226	265308	40665
V03	505705	505695	508600	503653	440135	392556	405149	47579
V04	155837	155795	156396	154006	129878	96881	101890	32997
V05	134025	134024	134655	130010	103304	83586	88888	19718
V06	140778	140775	141283	137818	111699	88064	95566	23635
V07	253249	253223	254493	246178	195996	162333	168732	33663
V08	257278	257277	258396	249598	198508	161048	168483	37460
V09	406923	406921	409115	389774	298514	250073	264639	48441
V10	276556	276497	278026	266712	202700	168201	178330	34499
U	551825	551420	448392					

The cDNA set was the most complex to process (Table 15) as the SISPA method introduces several sequence-related artefacts that must be addressed prior to sequence splitting, including truncated and tandem repeated SISPA barcodes within the sequences, incomplete adaptors and, in general, very short reads (because each SISPA barcode consist of 20 nt that must be removed and there are normally two per sample). For these sets, two scavenger sets of processes were run, the first to try to get some more sequences from the unidentified pool into the samples and the second for recovering some more sample sequences after dealing with the internal SISPA sequences.

Table 15. Summary of the cDNA set processing.

Pool	Raw	Derrep	Scavenger	Bad index	Adaptors	Sample	2nd Scav.	Adaptor remnants	Qual. Filter	Joined	JnS	Split
R1	1224979	1218556	1226980	1224025	1060429	V05	27149	7665	6438	5594	5876	844
						L08	704966	22356	7863	5997	8943	1866
						S07	252638	216368	174302	153970	158723	20332
						V06	47302	2177	1131	972	1232	159
R2	938075	919009	925047	917264	586233	S16	273161	261966	233592	215854	224507	17738
						L01	163731	32109	17119	15033	16569	2086
						L10	17635	4699	43	36	101	7
						C03	100814	87728	74642	67632	73796	7010
R3	979783	965861	972359	969111	756385	S06	225718	186217	174571	155815	161037	18756
						S13	204485	173002	160975	141812	148522	19163
						V08	129336	126619	122421	109094	110541	13327
						V02	143856	130384	118668	107277	114360	11391
R4	929357	918506	926835	924345	709202	L07	229724	199386	175644	156170	162496	19474
						S10	148536	85302	63845	56014	58490	7831
						L09	28731	19368	18251	16137	16548	2114
						S12	250256	175219	163710	151749	157982	11961
R5	1139241	1105784	1109447	1100895	804171	S03	98410	84057	65097	58426	62323	6671
						L06	83387	64974	28030	24438	26464	3592
						C02	338886	319379	298645	275630	280337	23015
						V10	167989	127246	117346	103183	105864	14163
R6	1014641	989880	991892	986521	664733	S09	57526	51376	36771	31939	33637	4832
						V01	75423	62536	57754	53181	54217	4573
						S04	188703	168077	161434	149929	153489	11505
						S11	249006	193553	182904	170118	173796	12786
R7	1194924	1143411	1145293	1129574	738534	L04	115882	109786	99841	93007	96699	6834
						C01	248836	228202	207385	184068	189944	23317
						C05	22207	20461	14988	13415	13932	1573

						V07	155481	128546	110908	102983	106509	7925
R8	1031301	1005892	1007917	1002033	941686	V09	131197	114814	112424	100045	101475	12379
						S02	358810	334492	308360	275567	280082	32793
						L05	238967	219992	206400	188437	191254	17963
						S08	121262	105060	90291	81126	82629	9165
R9	719354	706434	707129	704389	642440	L02	194970	180801	175043	162782	164440	12261
						V04	86514	70510	51222	46678	47815	4544
						C04	286219	257958	245174	222676	225961	22498
						L03	7926	6982	4605	3613	4272	992
R10	840006	827892	829252	825593	786500	S01	153437	138184	112677	102097	103608	10580
						S05	45445	41395	33457	30585	31201	2872
						S14	221069	187150	182960	178181	180195	4779
						S15	297184	197142	181246	170313	176861	10933

The resulting sequences were then clustered into four datasets: cDNA_JnS, cDNA_split, DNA_JnS, and DNA_split, containing 1014151, 401766, 4116958, 1397682 respectively. These were used to search in databases.

Databases

Datasets downloaded on August 8th, 2016 contained 460 genomes from archaea, as well as 532 other sequences from gene source (which may be other gene proteins in different strains or incomplete sequences; see Methods), 193 genomes from fungi and 53,534 other related sequences, as well as 64,342 bacterial genomes, as well as 13,536 other related sequences.

The viral database was far more complex, including 67 Archaeal viral sequences, 2,480 from phages, 211 from viroids, 4743 other viral sequences all from the European Nucleotide Archive, plus 2,124 phage genomes from the EBI's FTP. The NCBI sequences included 5,849 fna or ffn (see methods) as well as 7,182 gene source sequences from ASN.1. Finally, the INSD release contained 2,002,490. Everything was downloaded on November 11th, 2016. The PHAST prophage database had 23,330 sequences, most of them partial sequences.

The databases that were constructed had their taxonomy altered by creating additional categories for the unclassified rank labels (by adding "_n"; see Methods) so that unclassified items at an any given level did not mix with other unclassified items of different origin. The resulting filler labels expanded the number of observations as seen in Table 16 and the total number of records, size and total bp in Table 17. It is important to note that the species number remains unaltered. The number of changes were particularly notorious in viruses, which have more missing categories. However, this does not imply there are more species but rather that they will not be grouped together with other unclassified just because a rank is missing, making the taxonomy more robust to empty categories. In the case of viruses, the sequences were kept at the type level if available, and are thus far more separate records.

Table 16. Total number of different tags in the compiled databases.

Database	Phylum	Class	Order	Family	Genus	Species
ArchaeaDB_2017_01	5	13	24	39	114	363
BacteriaDB_2016_08_08	47	89	196	414	1,708	10,076
FungiDB_2016_08_08	8	27	59	122	189	354
PHASTDB_2017_01_10	10	17	43	84	172	477
Database	**Baltimore**	**Order**	**Family**	**Subfamily**	**Genus**	**Species/Types**
VirusDB_2016_11_16	1,343	1,350	1,446	1,473	1,990	15,629

Table 17. Database size, length and total records.

Database	nº items	Size	Length (bp)
Archaea_2017_02_01	25,169	1.3 G	1,301,137,567
BacteriaDB_2016_08_08	43,361,651	55 G	57,608,757,122
Plasmids	9,456	656 M	685,963,604
FungiDB_2016_08_08	175,875	5.4 G	5,795,494,963
VirusDB_2016_11_16	1,186,350	1.3 G	1,290,435,650
PHASTDB_2017_01_10	3,284	131 M	136,629,000

The databases were used for taxonomic identification for the Illumina clustered datasets cDNA and DNA using megablast, blastn, bowtie2 (in all permutations of databases and methods) and, for searching the VirusDB database, tblastx and filtering hits with multiple databases. The resulting contingency tables contained different types of organisms and very variable success of assignation within each set (Table 17) but clearly, the DNA set had more assignments. Intermediate steps in the process have no biological relevance since clustered sequences were employed instead of the raw sequences. In the end, each query was assigned a taxonomy based on the LCA of all the hits involved. Only taxonomic ranks that were congruent in all hits were included in the output and if more than one rank was found (i. e. different hits had different labels at that particular rank) then it was labelled as unidentified "-u" (see Methods) to ensure no cross contaminations would arise from multi database comparison. The resulting contingency tables containing all identified sequences are found in the online repository https://github.com/rodrigogarlop/PVL inside the raw tables folder.

In the cDNA dataset, a total of 23.32% of all usable sequences (after quality processing) remained unclassified as homology search produced no valid alignments with references from the different databases. In the DNA set, a total of 63.58% of the sequences remained unidentified.

Table 17. Totals/percentage of sequences that managed to get a taxonomy in the DNA and cDNA sets

Sample	cDNA Seqs	Hits found	%	DNA Seqs	Hits found	%
C01	213261	51898	24.34	221063	148946	67.38
C02	303352	104396	34.41	285861	261189	91.37
C03	80806	13526	16.74	361113	142293	39.40
C04	248459	72444	29.16	132696	70011	52.76
C05	15505	4045	26.09	282226	267554	94.80
L01	18655	6567	35.20	227098	79242	34.89
L02	176701	13111	7.42	379315	284090	74.90
L03	5264	3110	59.08	233359	198018	84.86
L04	103533	3448	3.33	139219	81371	58.45
L05	209217	77442	37.02	336752	287782	85.46
L06	30056	11227	37.35	403928	319683	79.14
L07	181970	46545	25.58	380915	95790	25.15
L08	10809	2766	25.59	293388	189770	64.68
L09	18662	6857	36.74	150880	56266	37.29
L10	108	7	6.48	498501	410757	82.40
S01	114188	35720	31.28	87649	64360	73.43
S02	312875	131553	42.05	514839	449064	87.22
S03	68994	12064	17.49	269647	86320	32.01
S04	164994	11121	6.74	280556	132050	47.07
S05	34073	15416	45.24	408704	341683	83.60
S06	179793	66757	37.13	376802	282196	74.89
S07	179055	73764	41.20	439736	328847	74.78
S08	91794	16126	17.57	279647	224149	80.15
S09	38469	19921	51.78	120680	82199	68.11
S10	66321	24575	37.05	309166	151843	49.11
S11	186582	31212	16.73	192354	91118	47.37
S12	169943	10547	6.21	343215	292233	85.15
S13	167685	20454	12.20	305792	268621	87.84
S14	184974	9227	4.99	376626	20040	5.32
S15	187794	17074	9.09	228456	159430	69.79
S16	242245	54442	22.47	191702	146361	76.35
V01	58790	1949	3.32	177324	93476	52.71
V02	125751	31903	25.37	305973	99650	32.57
V03	NA	NA	NA	452728	212817	47.01
V04	52359	17728	33.86	134887	27143	20.12
V05	6720	3017	44.90	108606	98518	90.71
V06	1391	495	35.59	119201	86024	72.17
V07	114434	9056	7.91	202395	134521	66.46
V08	123868	43023	34.73	205943	43562	21.15
V09	113854	2350	2.06	313080	285672	91.25
V10	120027	24631	20.52	212829	81195	38.15

Hit distribution in the cDNA and DNA sets was mainly driven by the percentage of bacteria and eukaryotes Figure 52. Eukaryotes make up 84.74% of the total hits in the DNA set and 40.47% in the cDNA set, mostly composed of human hits, which were removed from the rest of the analyses. On the other hand, bacteria comprise 10.55 and 53.93% of the total hits in the DNA and cDNA, respectively. In both cases, the percentage of viruses that were recovered is relatively low at 4.7% of sequences in the DNA set and 5.59% in the cDNA set. Homologs of eukaryotic viruses were more prevalent in the DNA set whereas prophage homologs were found in larger proportions in the cDNA set, congruent to the bacterial content. Archaea were virtually absent in both datasets, comprising less than 0.01 % in both cases.

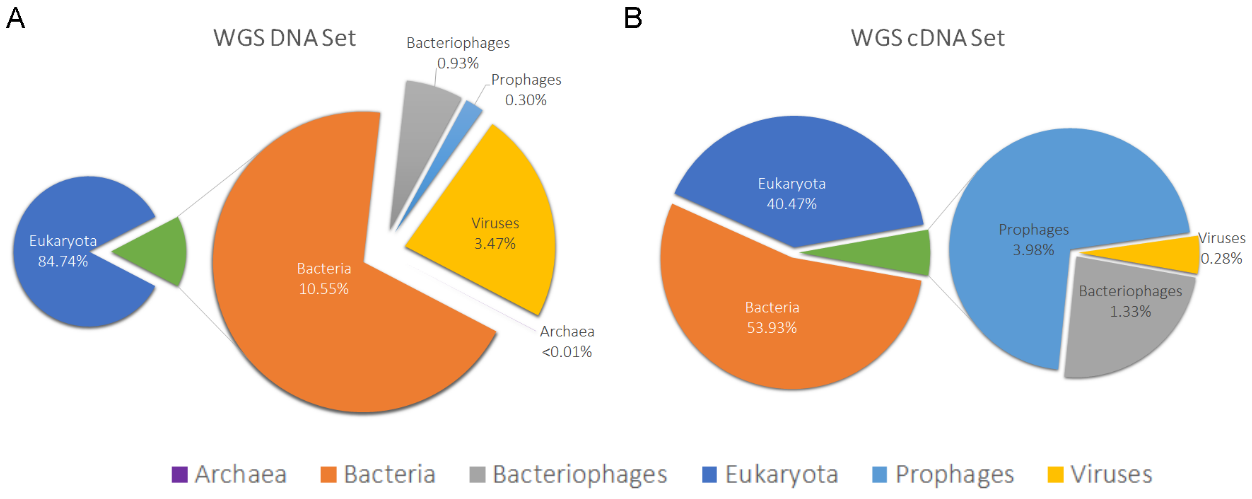


Figure 53. Taxonomic distribution of the type of organisms and viruses in the Illumina WGS sets. A) The DNA set had 4.7% viruses (all types but mostly eukaryotic viruses) and a higher prevalence of eukaryotic sequences. B) the cDNA set had a percentage of 5.59 % viruses and a higher prevalence of bacteriophages as well as a dominance of the bacterial sequences.

The different groups also displayed general compositional differences. The same trait for all cDNA is observed as bacteria is the most prevalent fraction in all. For instance, the OSCC set in the cDNA set had a larger percentage of human compared to the rest 49.83%, which may be due to the initial tissue disruption. In contrast, the PVL presented a larger percentage of prophages, at 3.94% (Figure 54). Yet, differences at this level are just broad observations as single individuals may be driving the compositional differences.

The OL group in the DNA set presented a larger proportion of bacteria (23.64%) whereas the PVL had a higher percentage of bacteriophages (3.12%) (Figure 55). Again, the proportion of human hits is the most prevalent feature of the dataset.

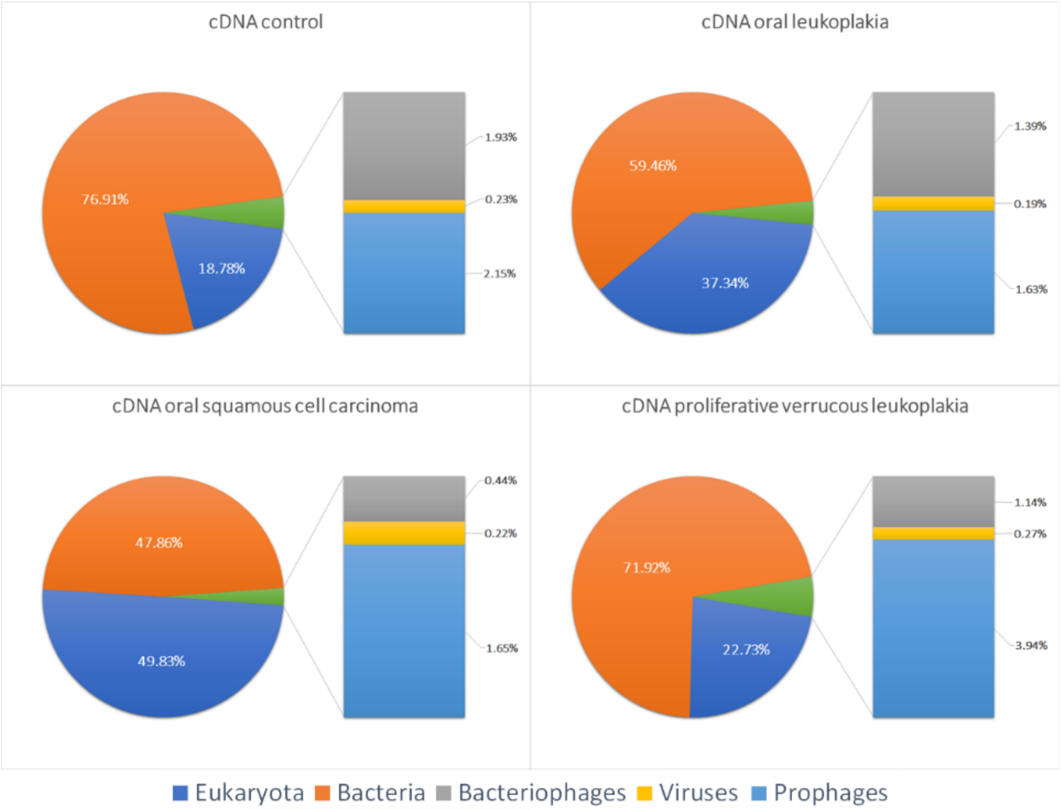


Figure 54. Percentage of hits in the cDNA set. The total abundances were scaled (relative numbers) to compare groups.

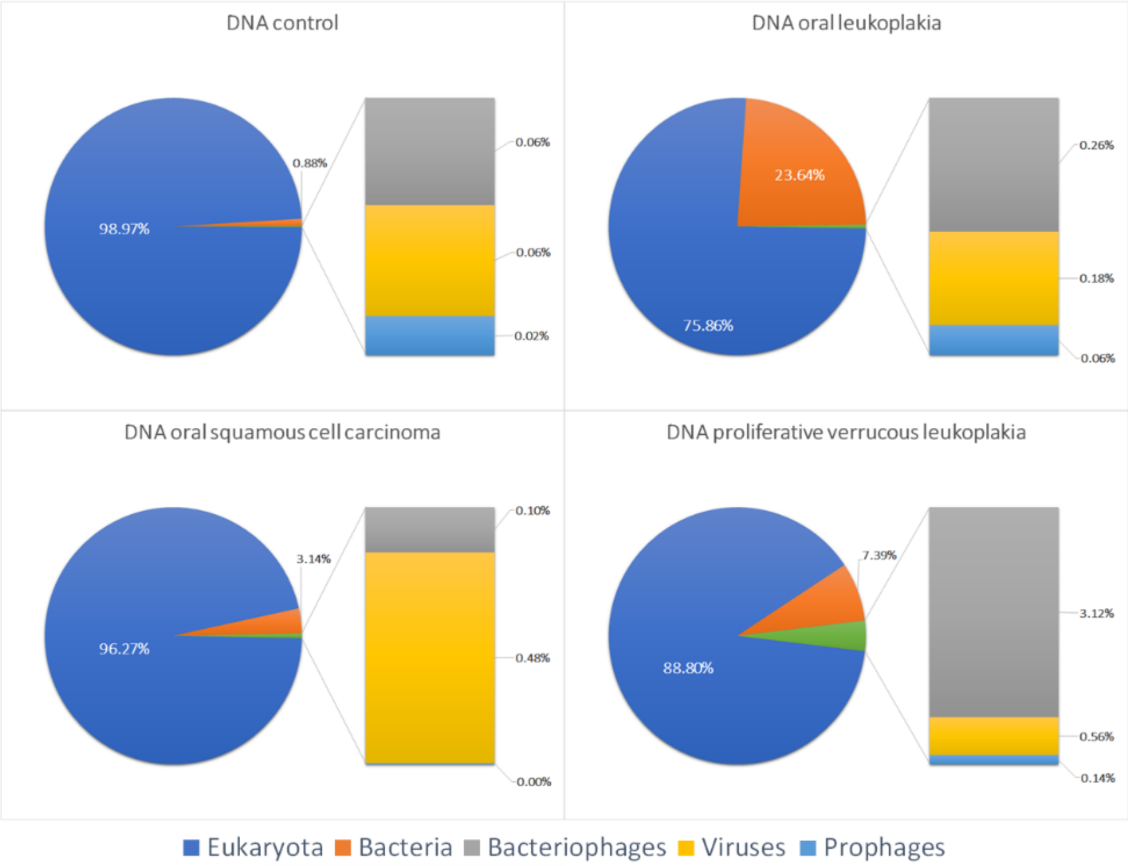


Figure 55. Percentage of hits in the DNA set. The total abundances were scaled (relative numbers) to compare between groups.

Compositional analysis.

Once the human samples have been removed from the Illumina datasets, the samples can be analysed for the total composition per sample by looking at the different fractions at different taxonomic levels. It is important to take into account that even with scaled data, differences among samples may not be directly comparable as the underlying abundance of the sample may be differential, even between samples belonging to the same groups. Also, finding a given organism or virus does not imply it was actually in the sample but, rather, that the LCA of the hit results pointed to a homologous sequence at that taxonomic level.

When examining the Baltimore classification, the most prevalent type of viruses in the DNA set were ssDNA viruses and they were most commonly found in OSCC samples. They also match the sampling effort. The dsDNA viruses were more evenly distributed (separated in bacteriophages and eukaryotic viruses. Control also showed a higher prevalence of retroviruses (Figure 56).

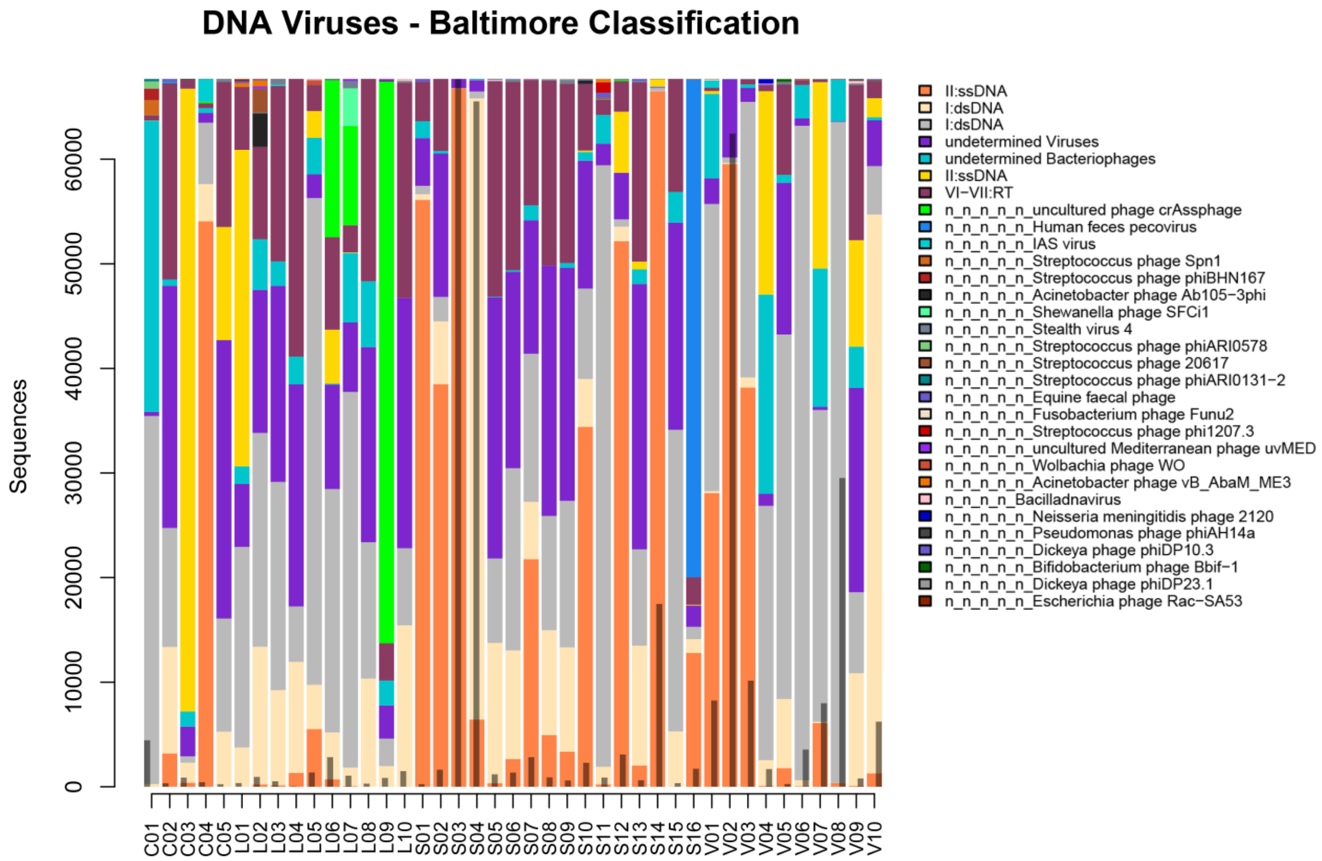


Figure 56. Composition of the DNA viral subset considering the Baltimore classification. For this and all the other compositional plots, the thin bars within broader ones represent the sampling effort of each of the samples and their scale is presented in the y axis. Broad bars represent the relative abundance of each OTU. Empty categories are marked with "n"s if there was no taxonomic information or "u"s when the hit was ambiguous at that level, pointing to the closest taxonomic rank available.

In the RNA set, for the Baltimore classification, the most prevalent ones were the negative ssRNA viruses, followed by the dsDNA (these were kept as there is a chance, although slim, of getting these during transcription) and retroviruses in part of the samples (Figure 57). It is important to notice the very low general abundance of the cDNA viruses in the samples.

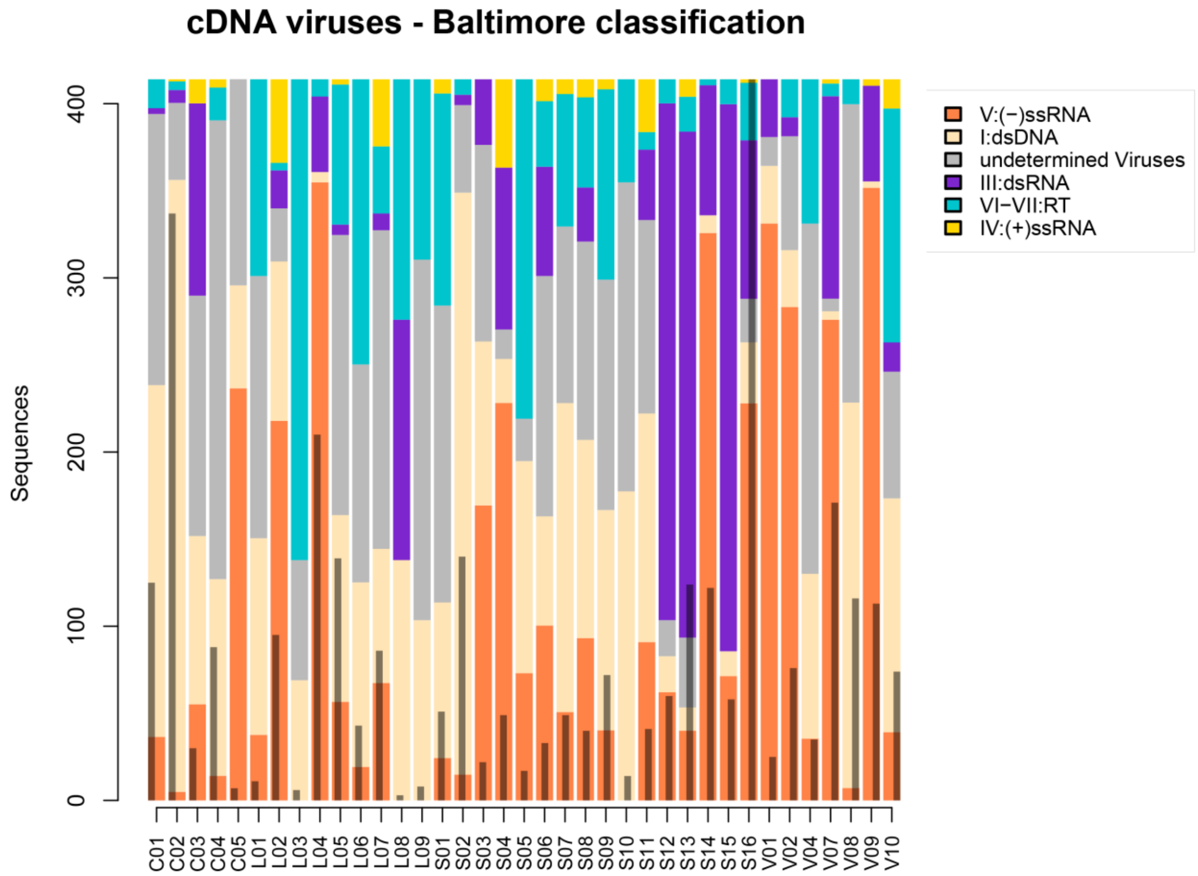


Figure 57. Composition and abundance of the cDNA viral subset considering the Baltimore classification.

The family level shed a light on the taxonomic composition of the sets. In the DNA set (Figure 58), family *Anelloviridae* was detected in OSCC and PVL sets. *Papillomaviridae* was in a low proportion but extended across all samples. The bacteriophage *Siphoviridae* family was also detected in most samples.

The corresponding comparison of the cDNA viral subset (Figure 59) was characterized by the presence of the *Arenaviridae* family. Sequences homologous to viruses from the *Herpesviridae* family were also found (kept as there was a possibility of getting the DNA viruses during transcription), the *Partiviridae* were also detected in some OSCC samples, as well as the *Orthomyxoviridae* family which was extended across most samples.

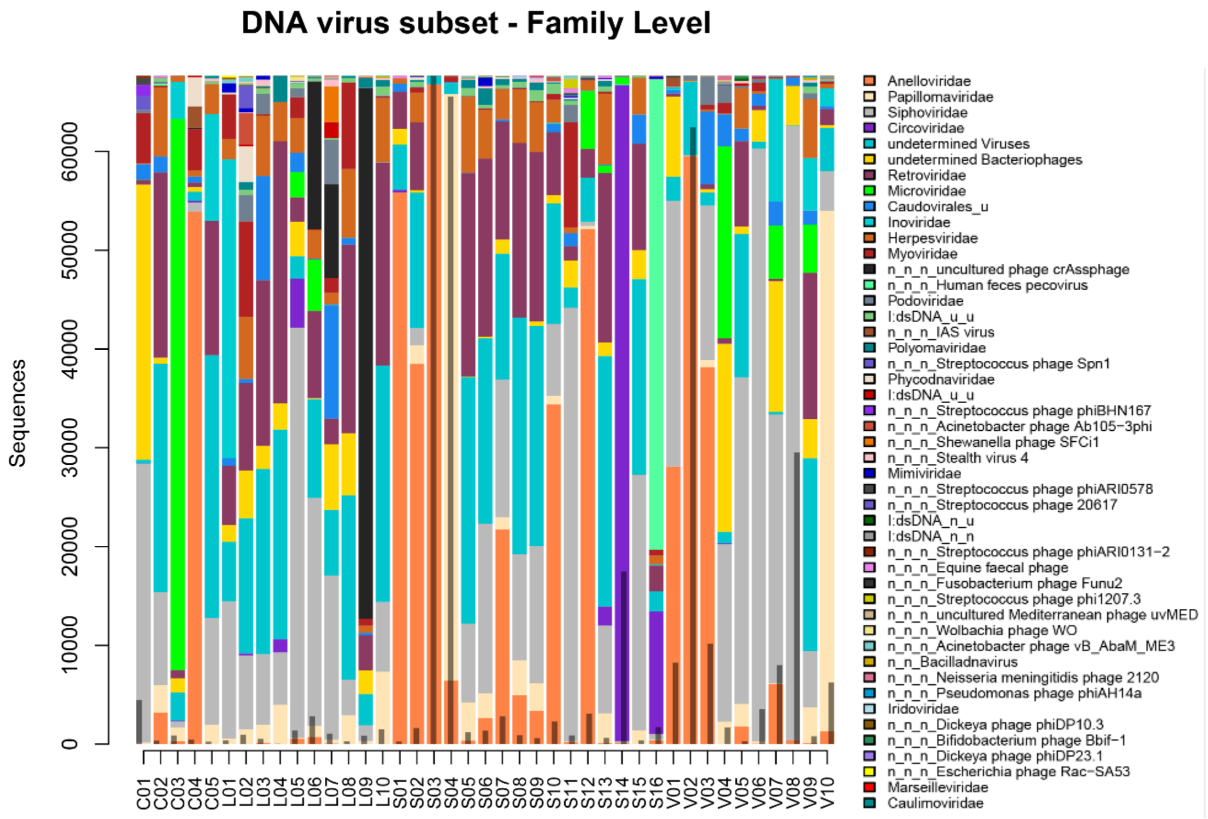


Figure 58. Composition and abundance of the DNA viral subset at the family level.

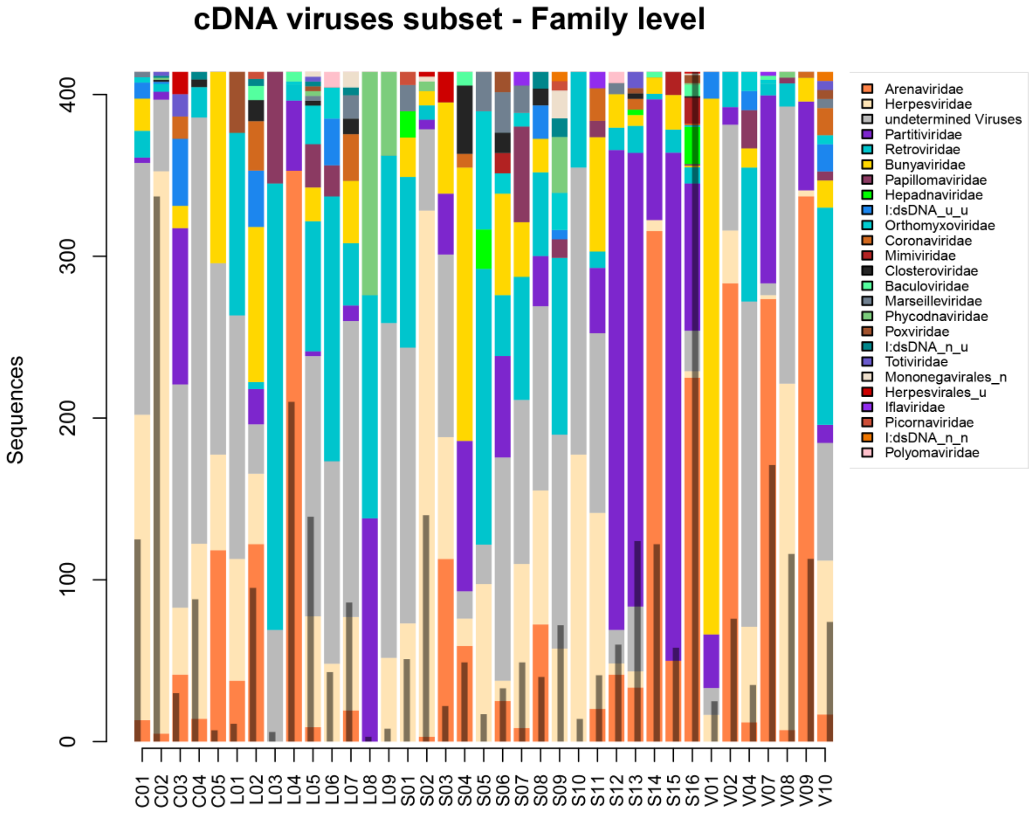


Figure 59. Composition and abundance of the cDNA viral subset at the family level.

At the genus level (Figure 60) it was clear most sequences with homology to *Anelloviridae* were in fact similar to the *Alphatorquevirus*, mainly composed of Torque teno viruses. Also, several unidentified genera from the *Papillomaviridae* Family were present mainly in the OL and OSCC samples. The species level (data for levels that are not show can be found in the online repository https://github.com/rodrigogarlop/PVL) revealed these to be sequences from various species of papillomaviruses, most of them with truncated resolution (hits were detected with more than one species) as well as some that were defined at that level such as the *Human papillomavirus type* 140, *Alphapapillomavirus* 9. Several bacteriophages were also detected, most of them similar to different *Streptococcus* phage.

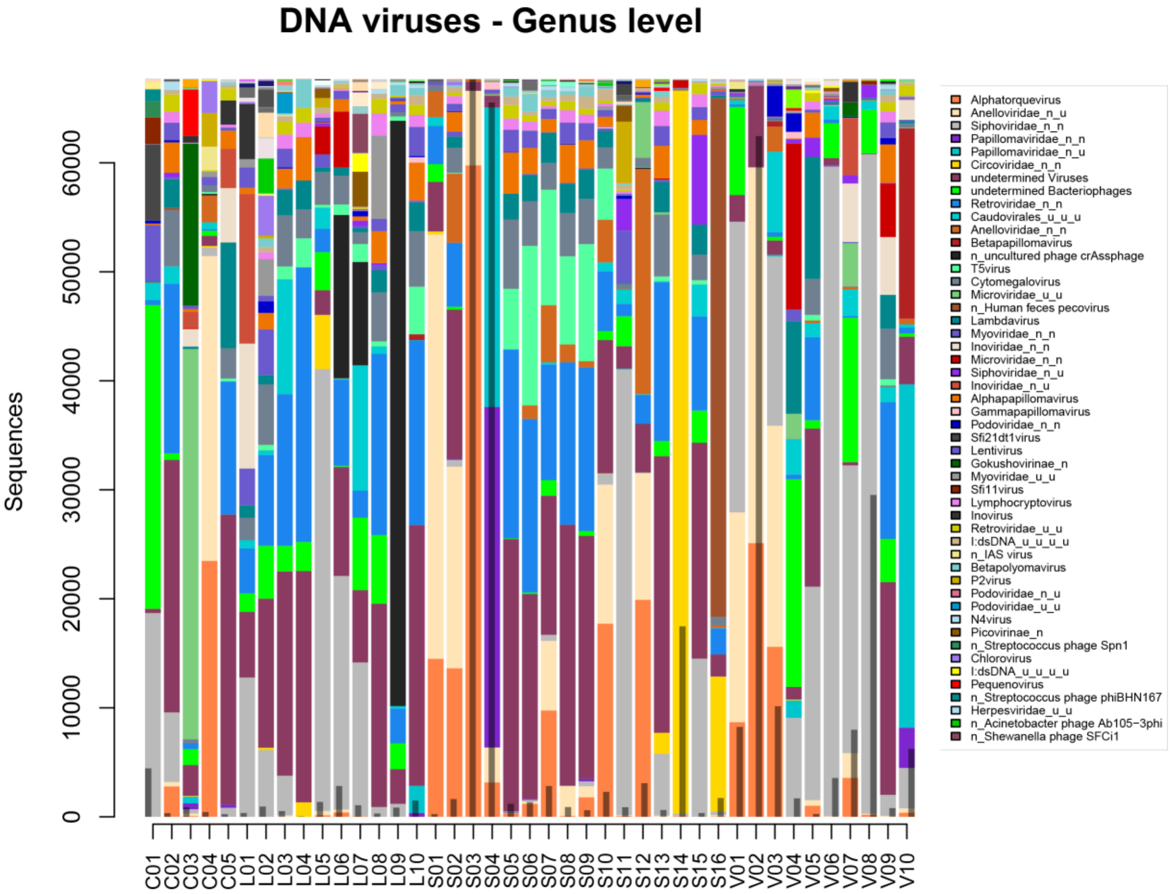


Figure 60. Composition and abundance of the DNA viral subset at the genus level.

The cDNA subset at the genus level (Figure 61) revealed *Mammarenavirus* as the most commonly detected in the samples, mostly found in PVL samples. The hypothetical DNA viruses in the set matched sequences from Cytomegalovirus, a mostly ubiquitous virus in humans while the rest remained unidentified. *Partitiviridae* genera were also found across all groups of samples.

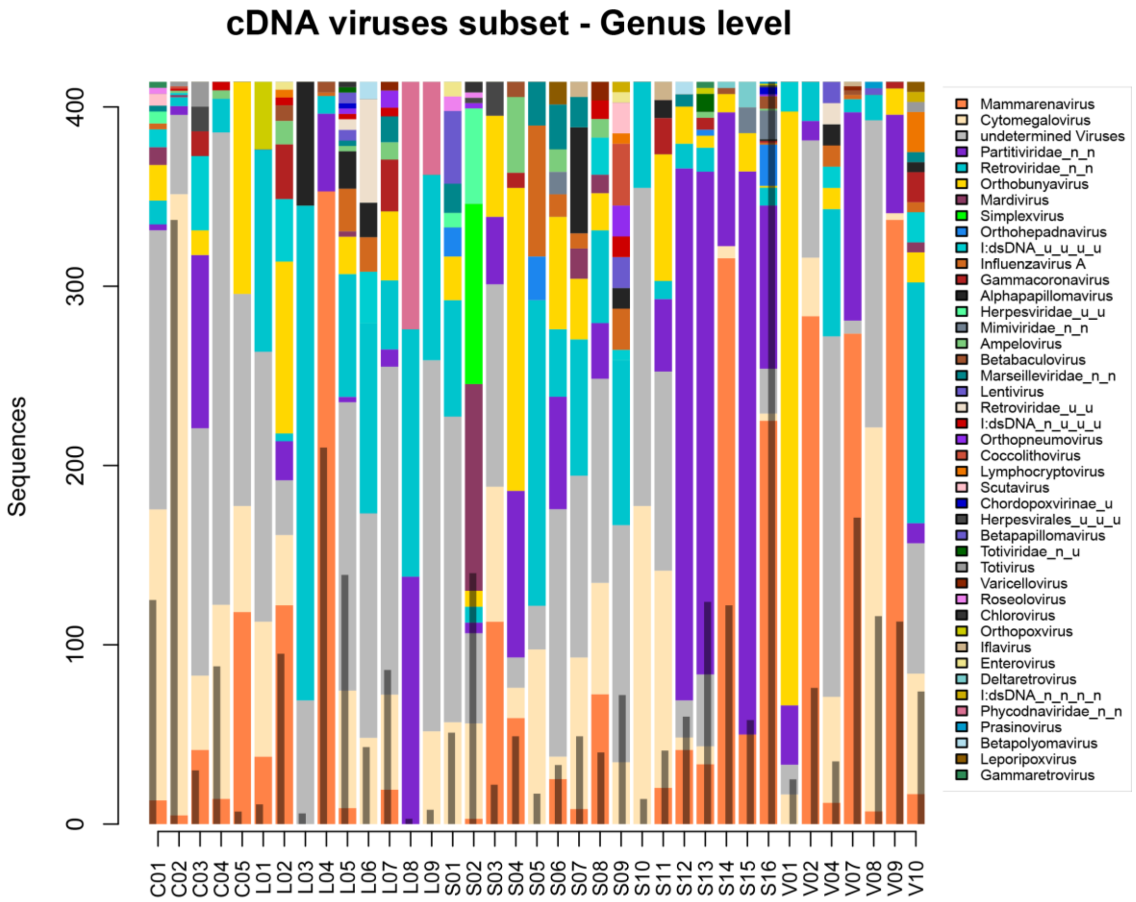


Figure 61. Composition and abundance of the cDNA viral subset at the genus level.

Fungi were not detected in all samples (or these were removed due to low number of fungal sequences). Only one control had these, and most were concentrated in OSCC samples and L03 (OL sample). At the genus level (Figure 62), the most extended fungi detected were those homologous to the *Candida* genus followed by *Malassezia*, a skin fungus, found in all groups in the study. A larger variety of fungi was detected within the RNA set (Figure 63), with *Malassezia* as the most prevalent and several species of yeasts including *Candida*. Yet, there were less sequences as well, which reduces the relevance of the observations.

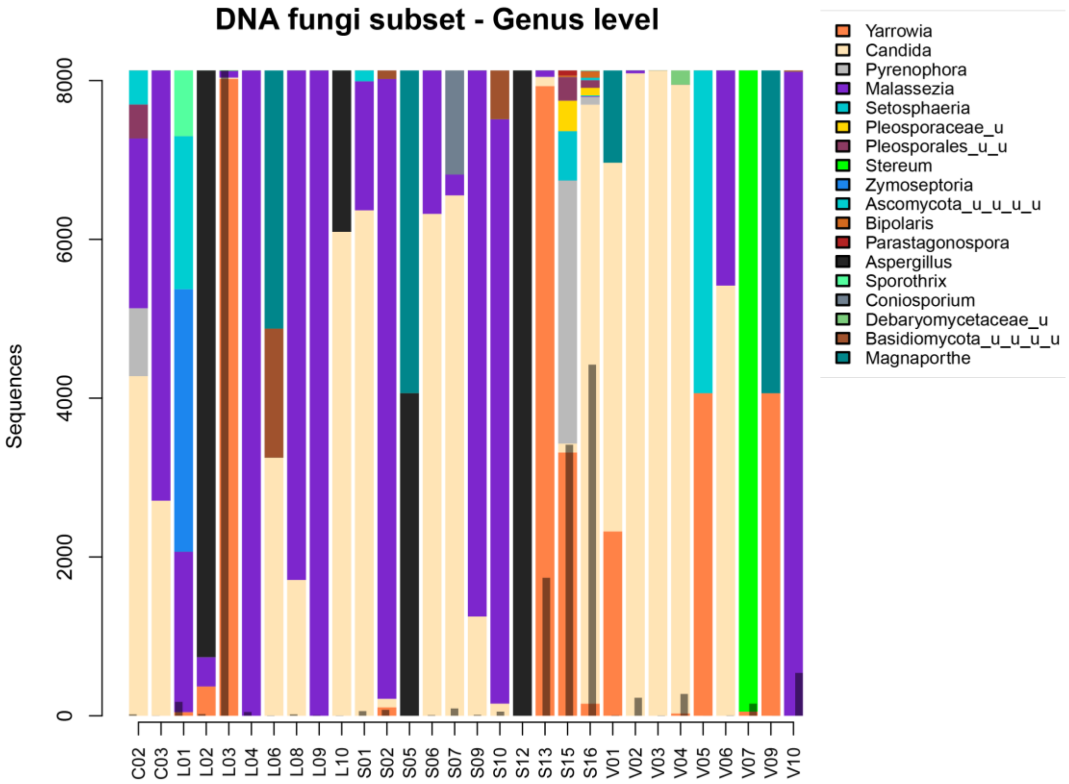


Figure 62. Composition and abundance of the DNA fungi subset at the genus level.

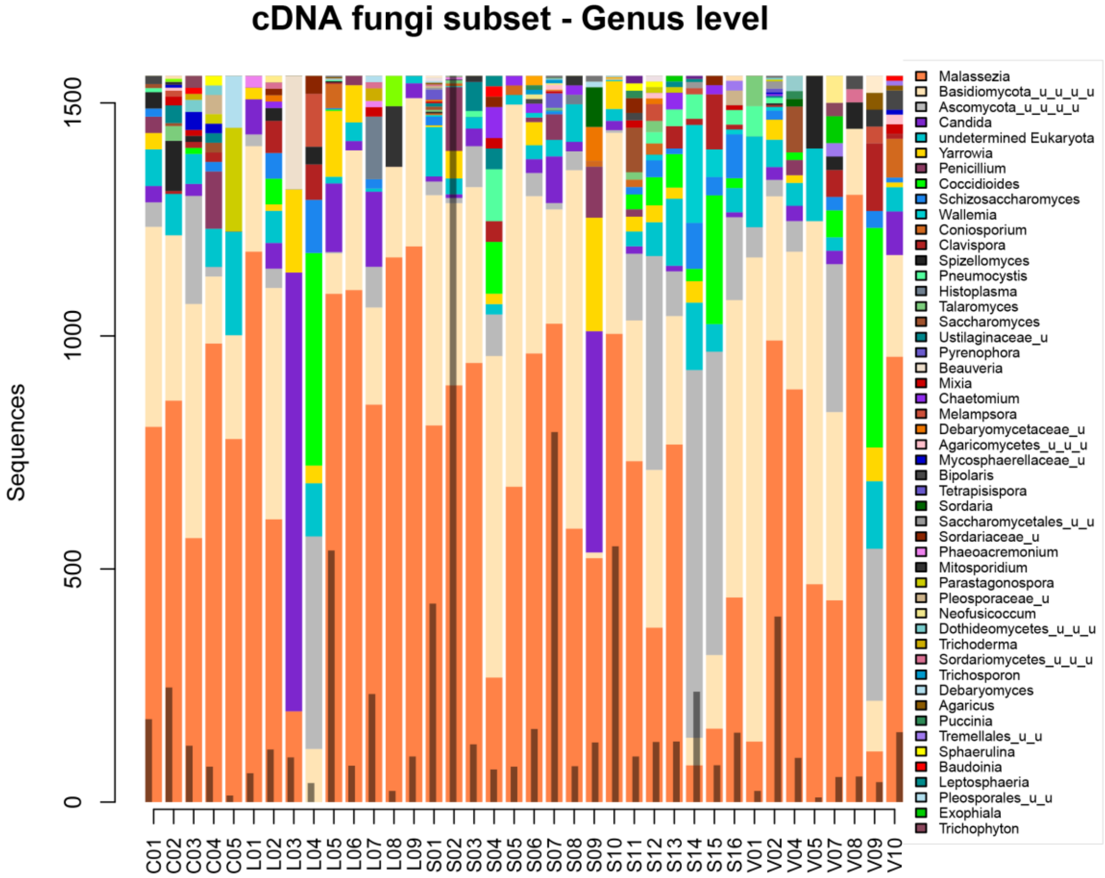


Figure 63. Composition and abundance of the DNA fungi subset at the genus level.

Prophage comparison was only carried out in the DNA set as these sequences are not normally transcribed to RNA unless they enter a lytic cycle. At the genus level (Figure 64), the most prevalent were a *Pseudomonas* prophage, a *Streptococcus* prophage and a *Haemophilus* prophage. These results were obtained from matches against sequences in the PHAST database that are in fact part of bacterial sequences that may have integrated viral DNA. It is important to note most samples were discarded in this sets and a few samples concentrate most of the observed counts

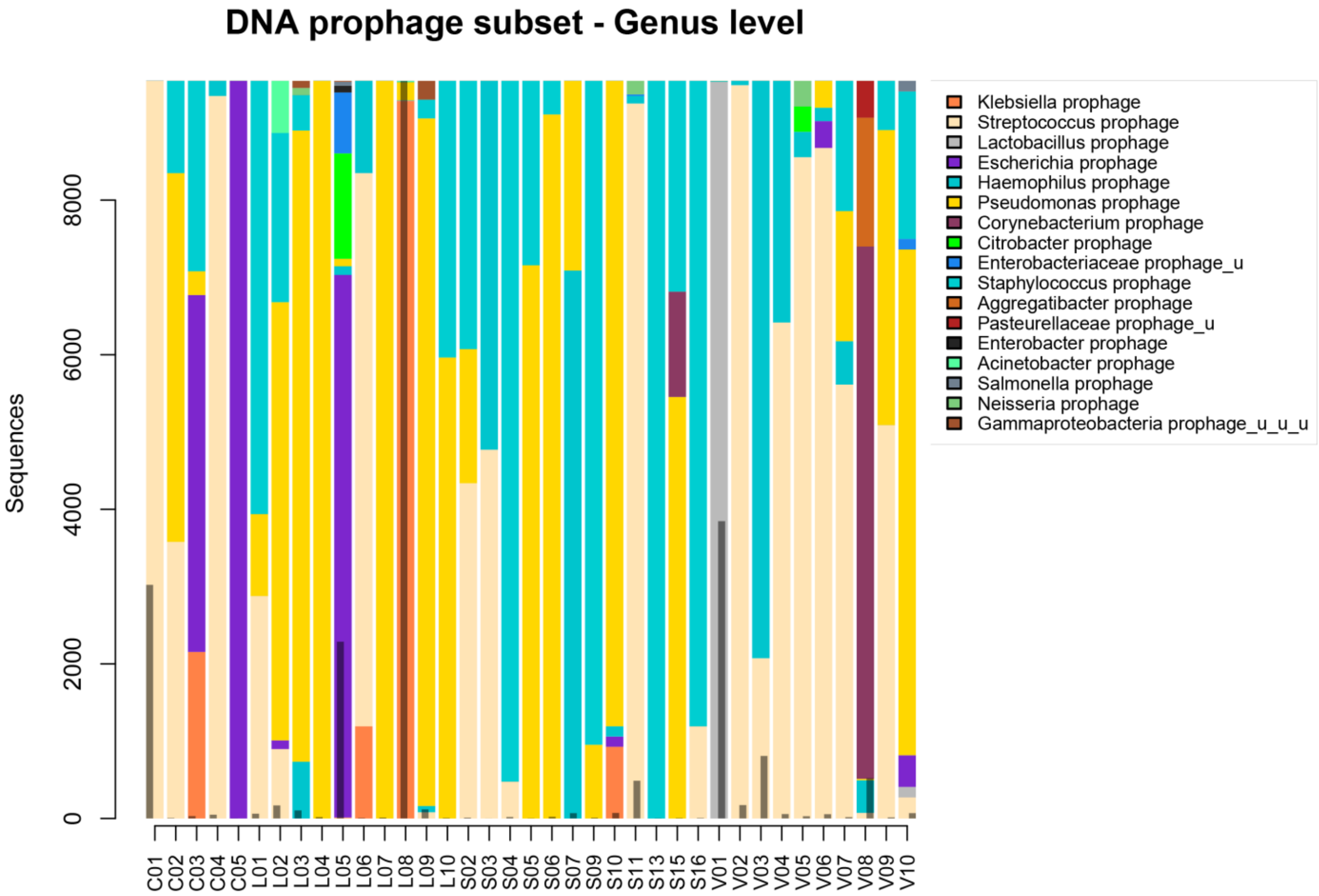


Figure 64. Composition and abundance of the prophages in DNA subset at the genus level.

The bacterial fraction of the Illumina sets presented a larger diversity as sequences were more abundant than for the other subsets. In the DNA set at the phylum level (data in repository), the most abundant group was Proteobacteria, followed by Actinobacteria and Firmicutes, distributed across the different samples and groups. At the genus level, *Pseudomonas*, *Streptococcus*, *Propionibacterium*, *Actinobacteria* and *Campylobacter* were some of the most prevalent species, distributed across all groups and samples (Figure 65).

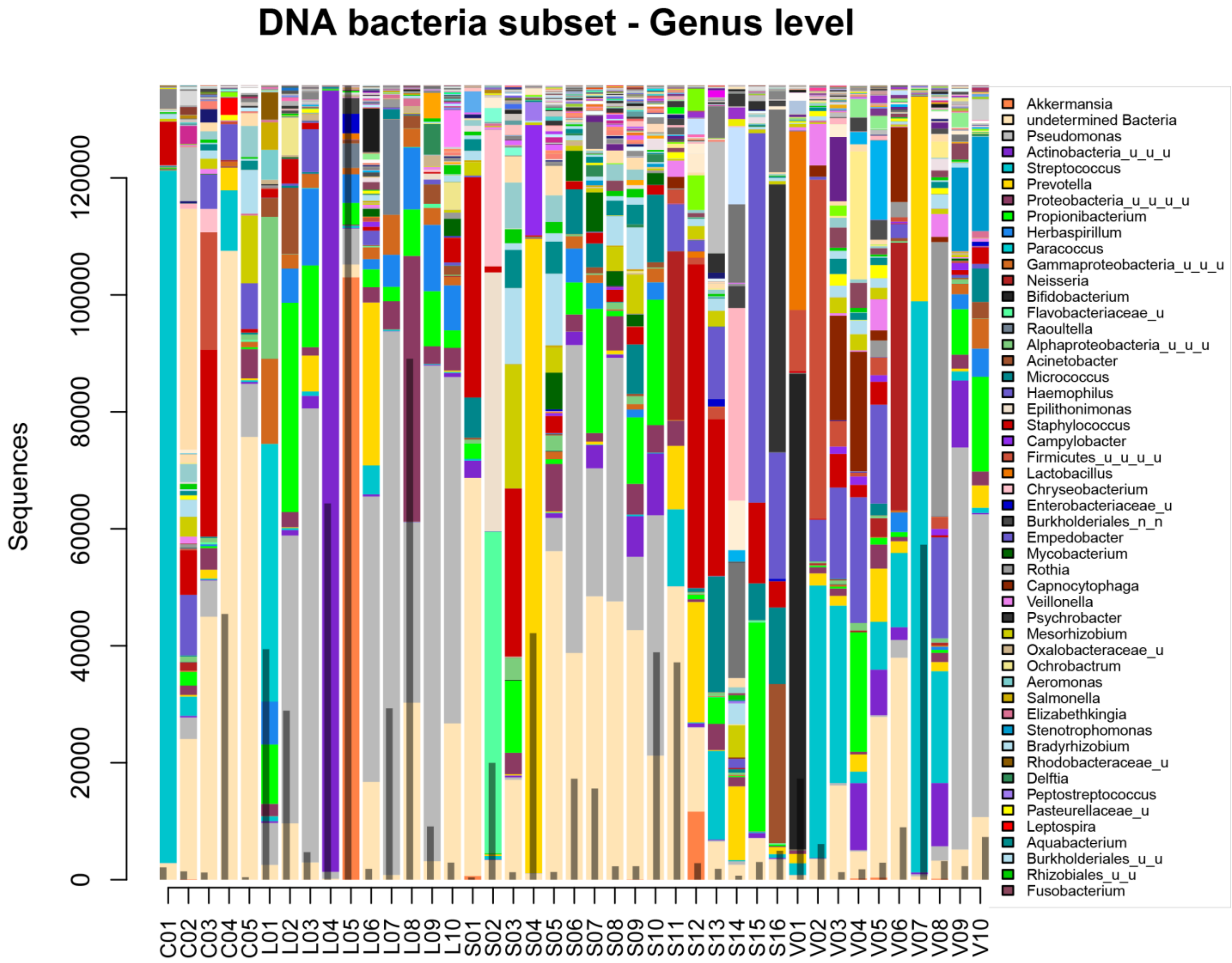


Figure 65. Composition and abundance of the bacteria in DNA subset at the genus level.

The cDNA dataset at the phylum level had a more uniform distribution across samples due to the large proportion of undetermined bacteria samples (hits that matched sequences from different phyla). This fraction was interesting, however, as it may represent the active part of the bacteria although it is mostly populated by 16S rRNA sequences. The most abundant fraction of the set was that of Actinobacteria, followed by Proteobacteria and Firmicutes. *Propionibacterium*, *Corynebacterium* and *Pseudomonas* were commonly detected at the genus levels (Figure 66).

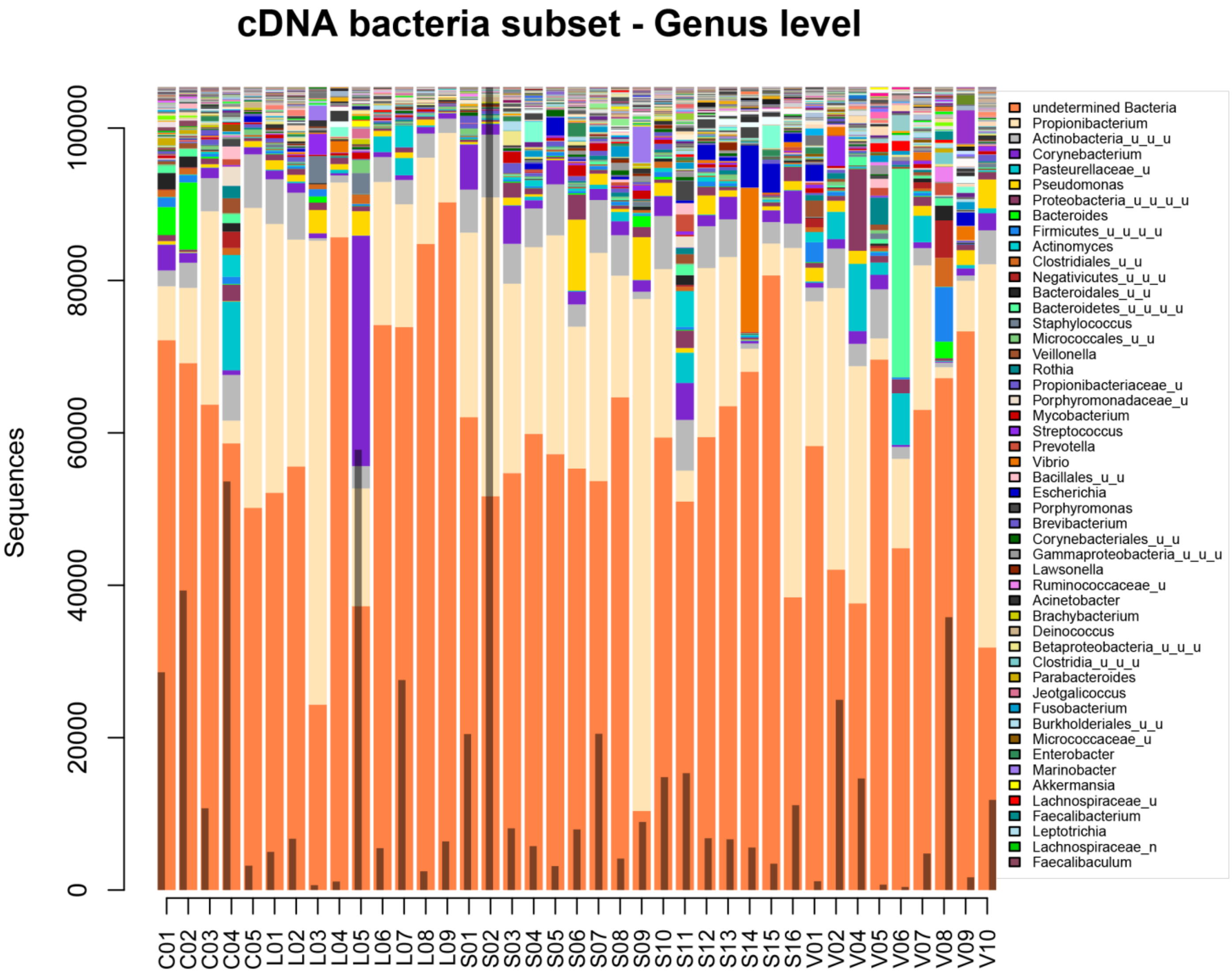


Figure 66. Composition and abundance of the bacteria in cDNA subset at the genus level.

Finally, the 16S profiles showed a somewhat distinct conformation at the phylum level, with a marked prevalence of Firmicutes, followed by Bacteroides, Fusobacteria, Proteobacteria and Actinobacteria. Minority groups included Spirochetes and Synergistetes.

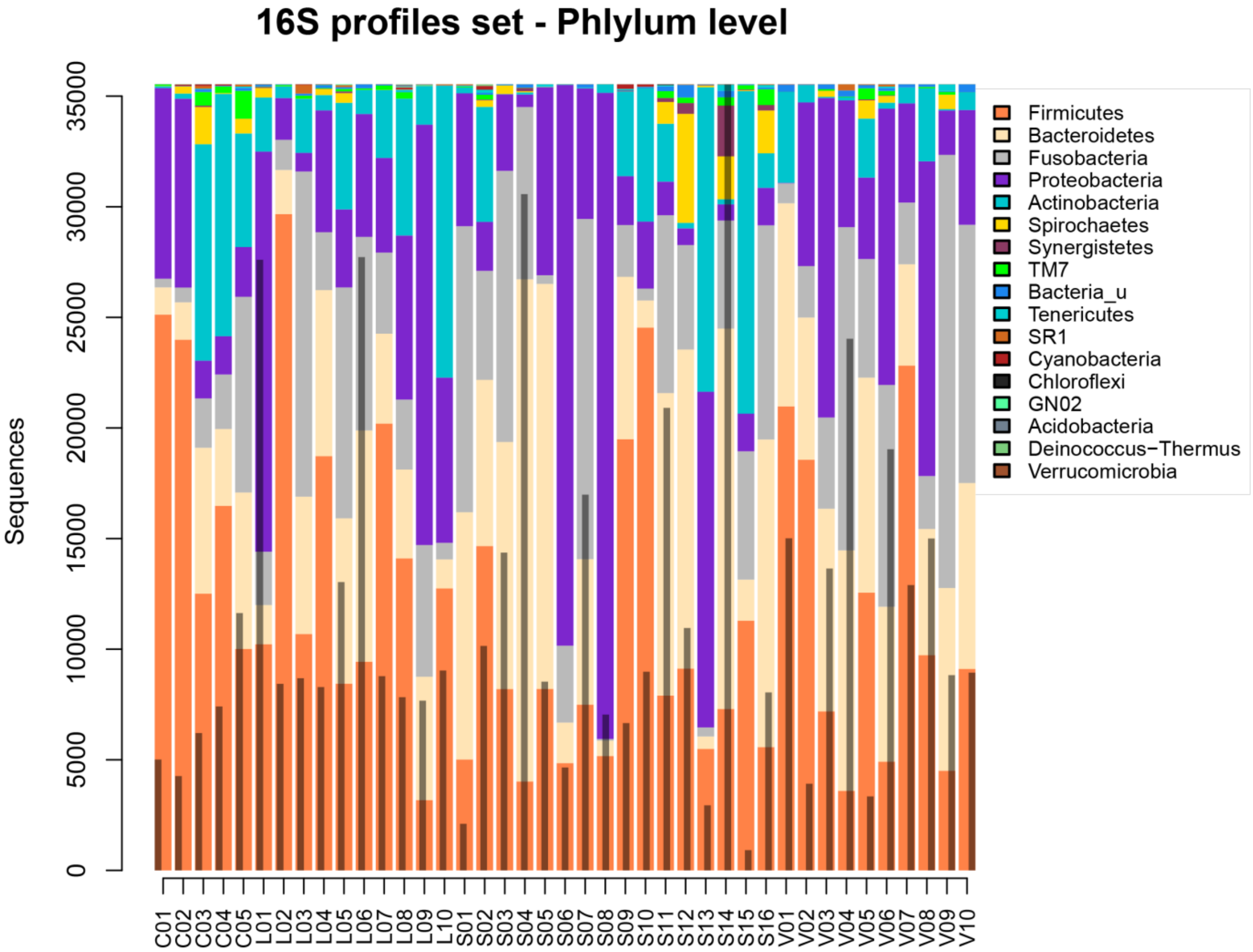


Figure 67. Composition and abundance of the bacteria in the 16S set at the phylum level.

The composition of the 16S profiles at the genus level (Figure 68) was far more diverse, with matching streptococci as the more prevalent in average but differentially distributed with no group pattern. Other genera occurring in all groups include *Haemophilus*, *Leptotrichia*, *Veillonella* and *Rothia*.

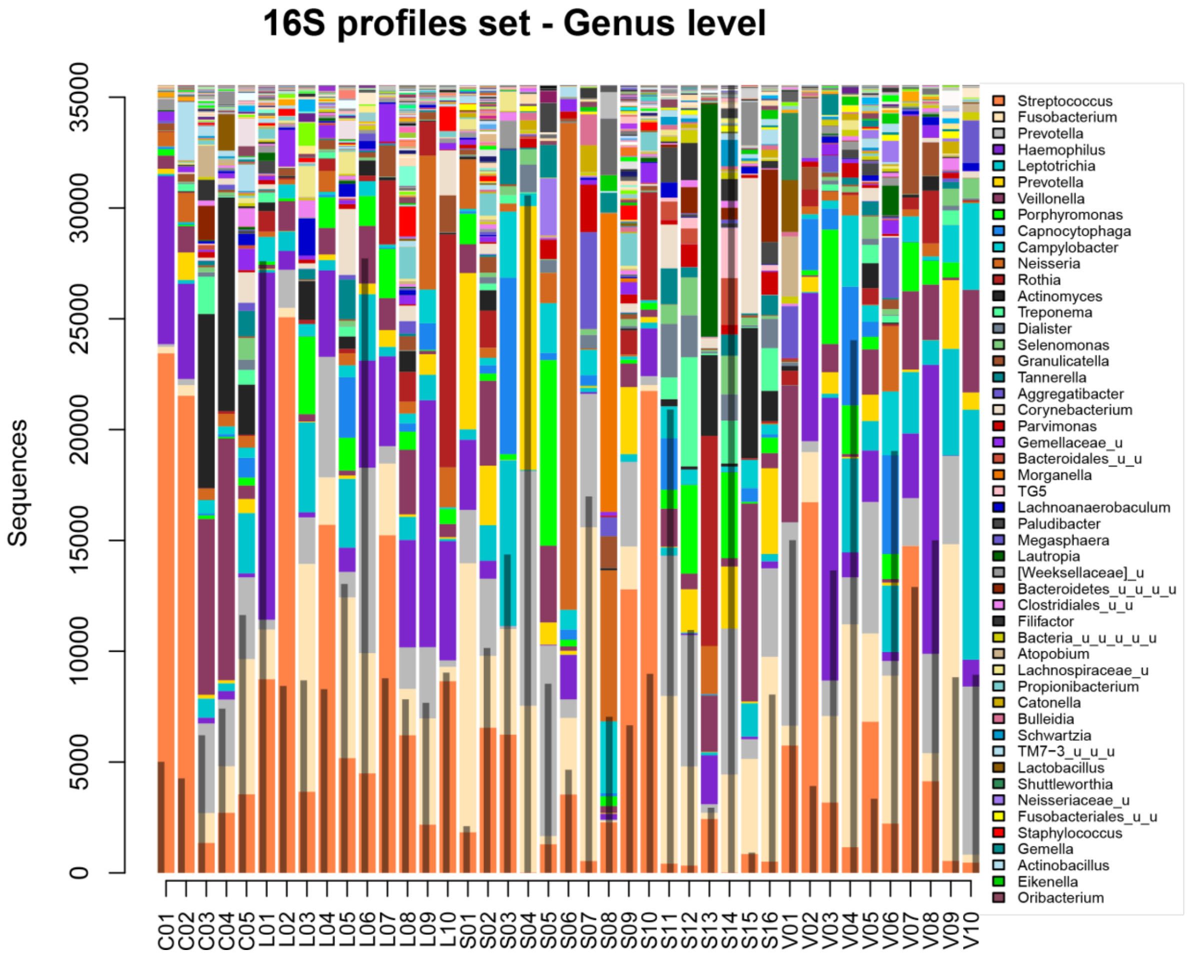


Figure 68. Composition and abundance of the bacteria in 16S set at the genus level.

No evident pattern of distribution was detected in the samples as most of the mentioned markers were found across all species. Depth of sequencing was markedly uneven, mainly in both Illumina sets. As a result, samples cannot be directly compared with only relative abundances.

Alpha diversity and rarefactions

Rarefaction curves (Figure 69) were calculated for the total number of species at increasingly larger resampling iterations. This was carried out to assess whether the sampling effort of each sample had been enough to recover most OTUs that may be present. It is expected that a plateau is reached once no new species is detected by increasing the number of total sequences. The plots are used to compare sampling in each set but not between them as methodologies were distinct.

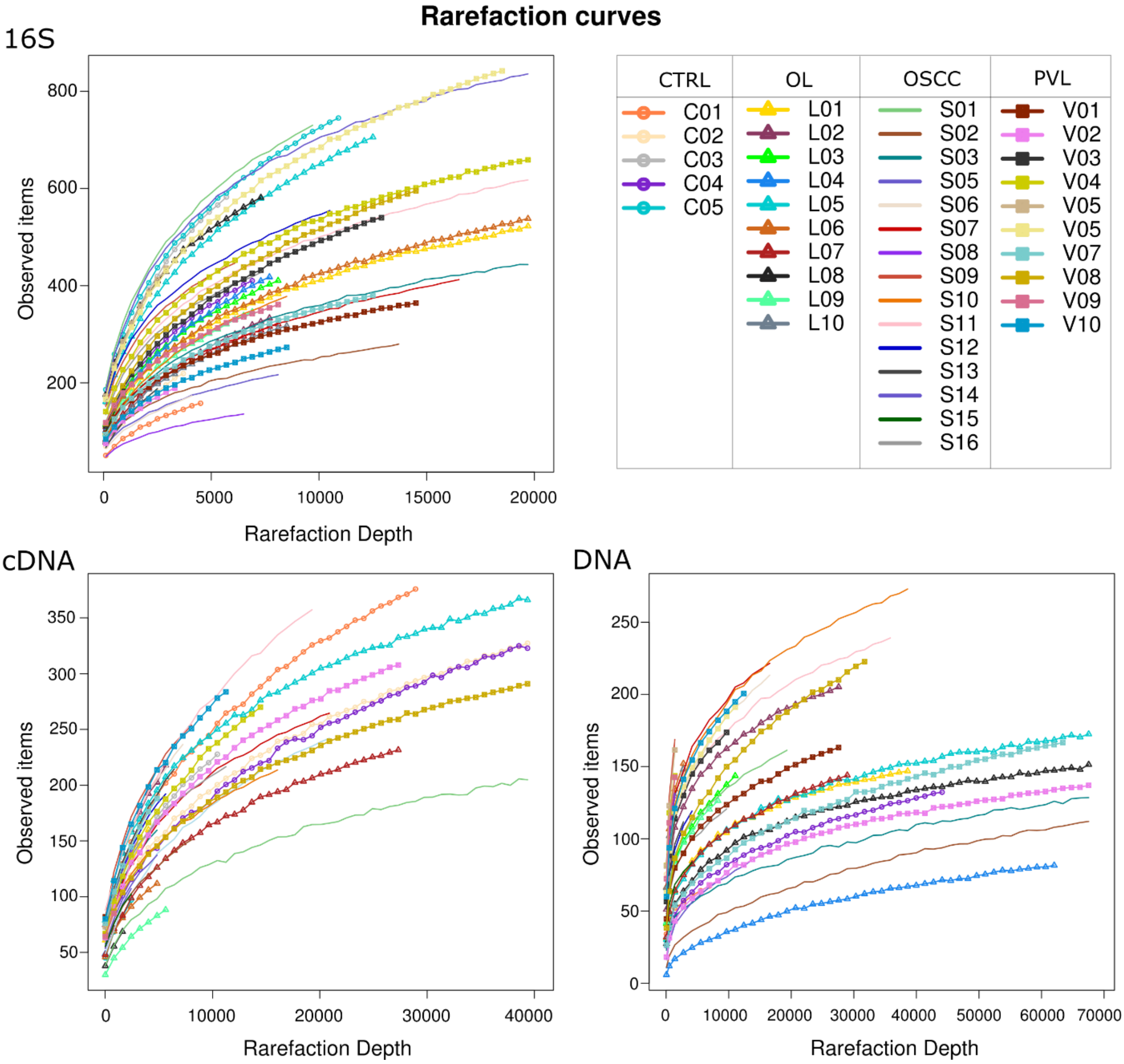


Figure 69. Rarefaction plots for observed number of items. In the case of 16S, the items refer to OTUs, whereas the Illumina sets used species counts.

Samples in the 16S rarefaction were the most saturated, followed by the DNA set and then the cDNA set (Figure 69). The inconsistency in the cDNA total number of sequences per sample was reflected in several samples having just present at the first few steps and a high slope in the truncated curves. The cDNA set, however, presented a higher variability as well since samples with the largest number of observed items were around 300. Group patterns were not evident from the graphs. Yet, due to the different methodology of the 16S profiles, the numbers are not directly comparable to the Illumina sets, as the former refers to OTUs, rather than species counts, even if set at 97% identity.

To assess variation between the samples, the Chao 1 richness estimator and the Shannon′s entropy indexes were evaluated in 1000 resampling iteration (rarefactions) on a depth of 3,000 (Figure 70). Samples having fewer than these number of items were dropped in favour of making a better comparison (the cDNA set was the most affected). Again, this is used to compare samples within sets but not between different methodologies. The singleton proportion reported with Chao1 was lower in the OSCC samples of the 16S with respect with other groups and higher in the PVL samples of the DNA set.

Further analysis of group differences provided no significant adjusted q-values (adjusted p-values <0.05) from the different comparison when non-parametric t-tests with 1,000 Monte Carlo iterations (FDR correction to compensate for multiple testing) were run with all category permutations, including pathology group, gender, age group and sample location. The statistical test results, along with the boxplot corrections for the different categories compared were deposited in the online repository under the name results _comparisons_alpha.txt.

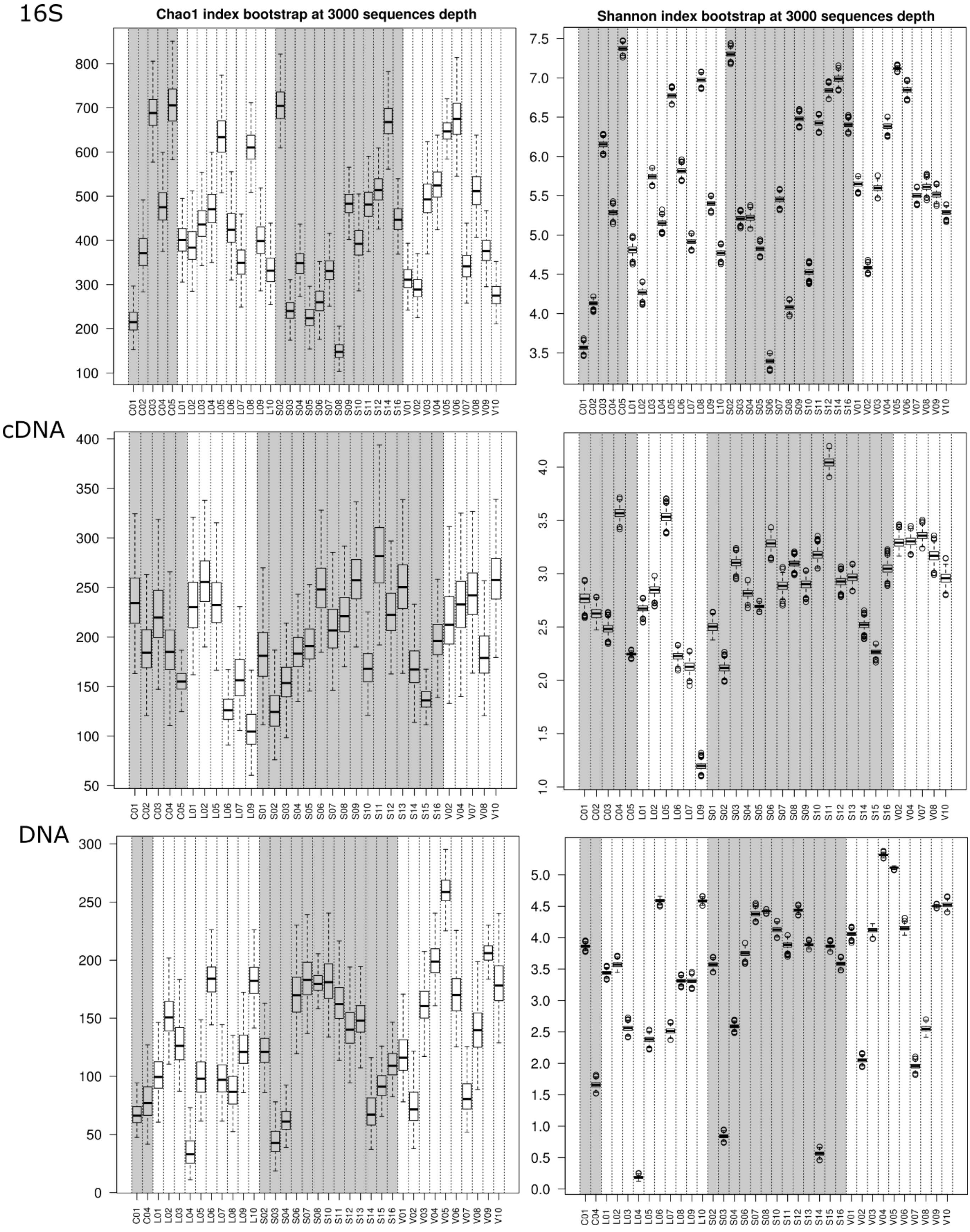


Figure 70. Alpha rarefaction index comparisons for the 16S, cDNA and DNA sets. A bootstrap consisting in 1,000 rarefactions was carried out at a depth of 3,000 sequences to compare the Chao1 and Shannon

index across the different samples. Not all samples are shown as some had fewer than 3,000 sequences and were discarded during the rarefaction step. Boxplots represent the 25 to 75 quantiles. Outliers were removed from the Chao1 plot to reduce image saturation.

Beta diversity

To compare the differential abundance of each sample in the datasets, variation was assessed with PCoA using Bray-Curtis dissimilarity index matrices whereas absence/presence was calculated using the Jaccard Index, both after sample normalization (or rarefaction in the case of viral sets), to create linear combinations (multidimensional scaling) based on composition that better explain the data. Between species variation was not clearly differentiated by group, as seen in the plots (Figure 71) except for the resulting transformation of the DNA-non-vir CSS normalized set (Figure 71E) which produced a separation of the OL, PVL and OSCC groups and was statistically significant for sample grouping by pathology category (adonis p-value < 0.001). In general, linear combinations built from the Jaccard index matrices explained a smaller percentage of the variation than the ones built with the Bray-Curtis metric and only produced a better cluster separation for the 16S profiles set (adonis p-value < 0.001; Figure 71B). The effect of the sequencing depth over the ordination method was evaluated by plotting a custom axe containing the depth (no separation by depth was detected). The PCoA for the cDNA viral dataset showed a marked separation by the fist axes which had no statistically significant match with any of the known categories (meaning it is improbable that any of them caused this divide.

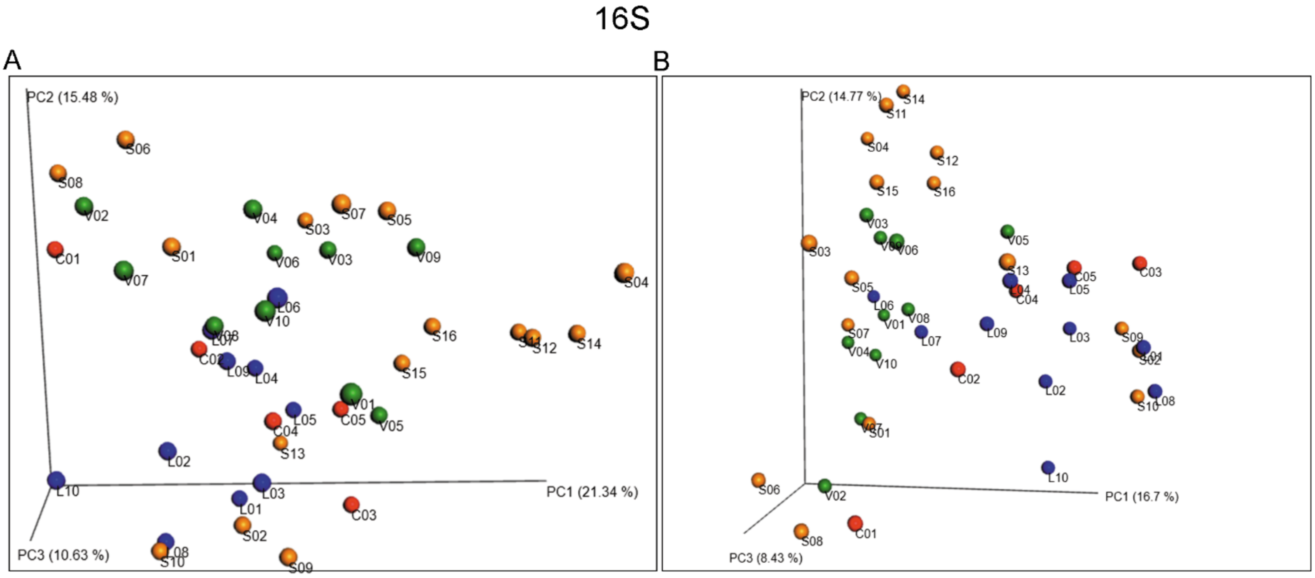


Figure 71 - Part 1. PCoA plots for evaluating beta diversity. The first three new axes were plotted in three-dimensional plots from the results of PCoA. A) Plot constructed from the Bray-Curtis dissimilarity using the 16S profiles sets (32.45% of variation explained by the tree fist axes. B) PCoA from binary Jaccard distance matrix using the 16S profiles dataset (39.9% of explained).

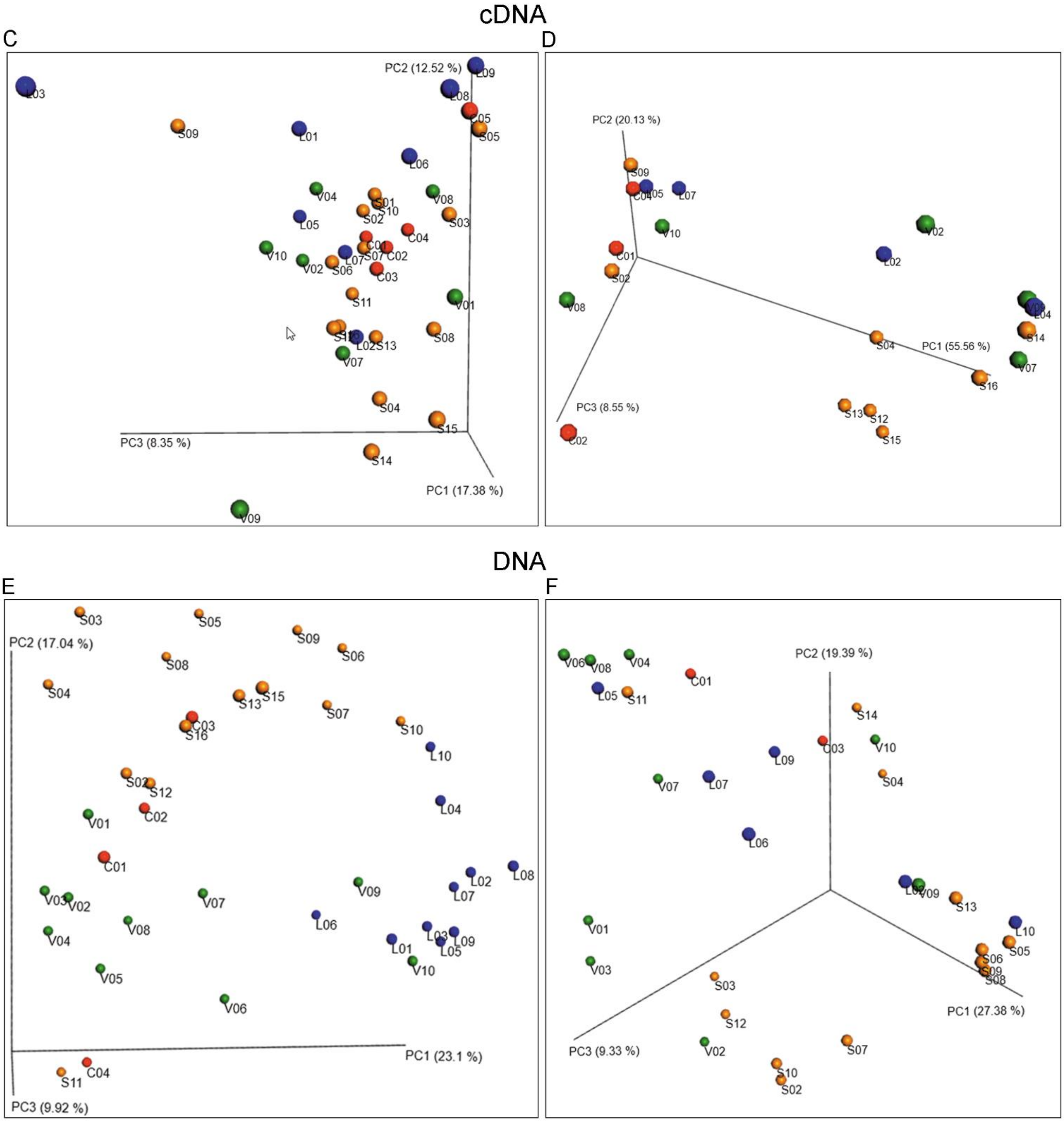


Figure 71 - Part 2. PCoA plots for evaluating beta diversity. C) PCoA from the Bray-Curtis matrix with the cDNA non-vir dataset (38.25% explained). D) PCoA from Bray-Curtis matrix using the cDNA viral dataset. Several samples were dropped during rarefaction (84.24% explained). E) PCoA from Bray-Curtis using the DNA non-vir dataset (50.06% explained). Group separation is achieved by the combination of the species composition. F) PCoA from Bray-Curtis using the DNA viral dataset (56.1% variation explained).

In the case of the cDNA non-viral subset, groups identified as well-defined in the PCoA were analysed by comparing distance matrix compartments formed for each group (Figure 72). The paired distances were evaluated using a two-sided t-test with Bonferroni correction to evaluate significance. Those with adjusted p-values < 0.5 include the All_between and the OL vs. OL (as well as the OSCC vs. OSCC) distances, meaning the groups OL and OSCC are better defined. The All-within and All between distances, as there is larger variation between the groups than within the group. Also, most combinations of the OL vs. OL and other groups of distances (the same as for OSCC and PVL). The diversity matrix for the cDNA viral dataset was also evaluated statistically but no significance was found for any of the groups. The 16S set using the Jaccard had differences between groups that were statistically significant, supporting the separation of the OL and OSCC groups.

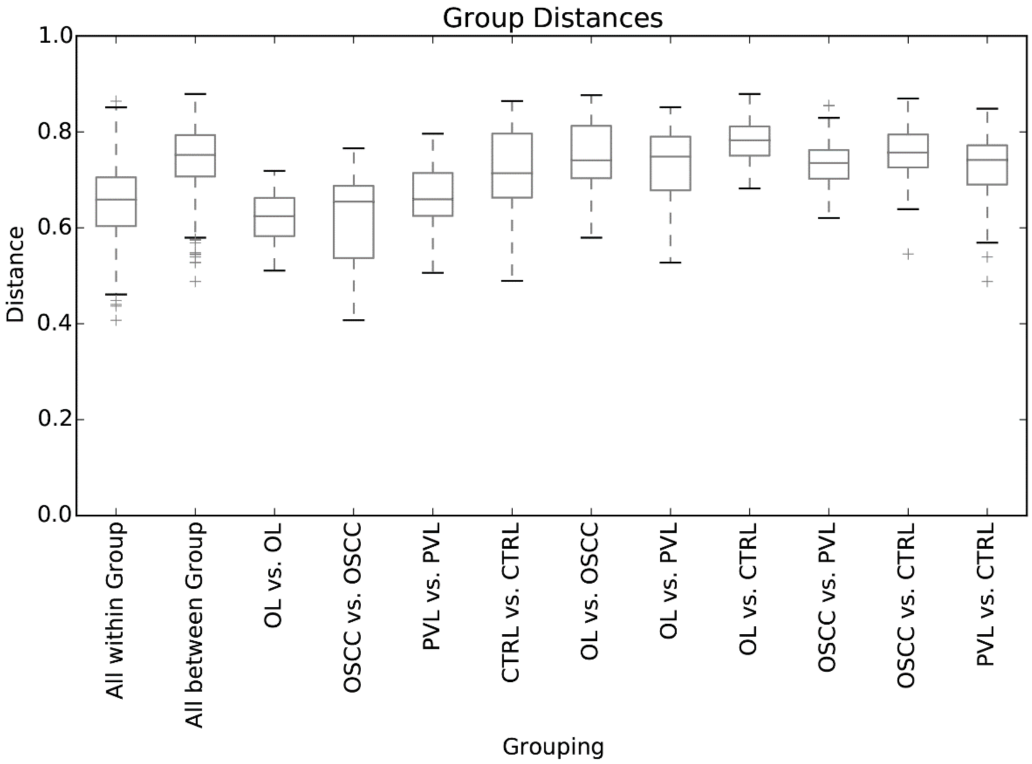


Figure 72. Boxplot of different compartments of the Bray-Curtis matrix calculated from the DNA non-vir subset composition in the beta analyses.

Heatmaps

In order to evaluate the differences producing the clusters in the Beta diversity analyses yielding significantly different groups, heatmaps were produced for all datasets at all taxonomic levels for the three datasets (16S, and, cDNA and DNA). Contrary to the previous composition plots, these allow for direct comparison of the contents of a group as a scaling and log transformed is used and the source contingency table is filtered (strict tables). In this case, they were ordered by pathologic group. In the cDNA heatmap displaying the Baltimore classification, the only interesting remark is a higher prevalence of (-) ssRNA viruses in some of the OL samples and undetermined viruses (Figure 73).

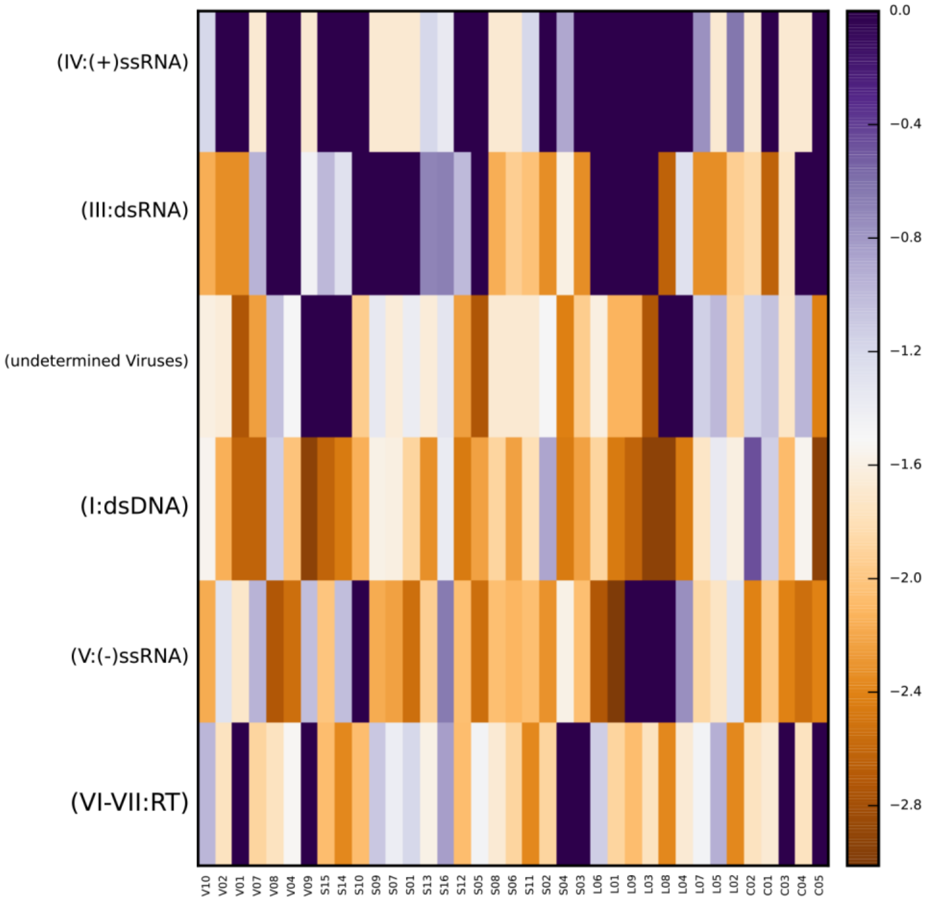


Figure 73. cDNA viral subset – Heatmap using Baltimore classification.

The family level heatmap showed a higher prevalence of the *Partiviridae*, *Bunyaviridae* and *Arenaviridae* families OSCC and OL (Figure 74).

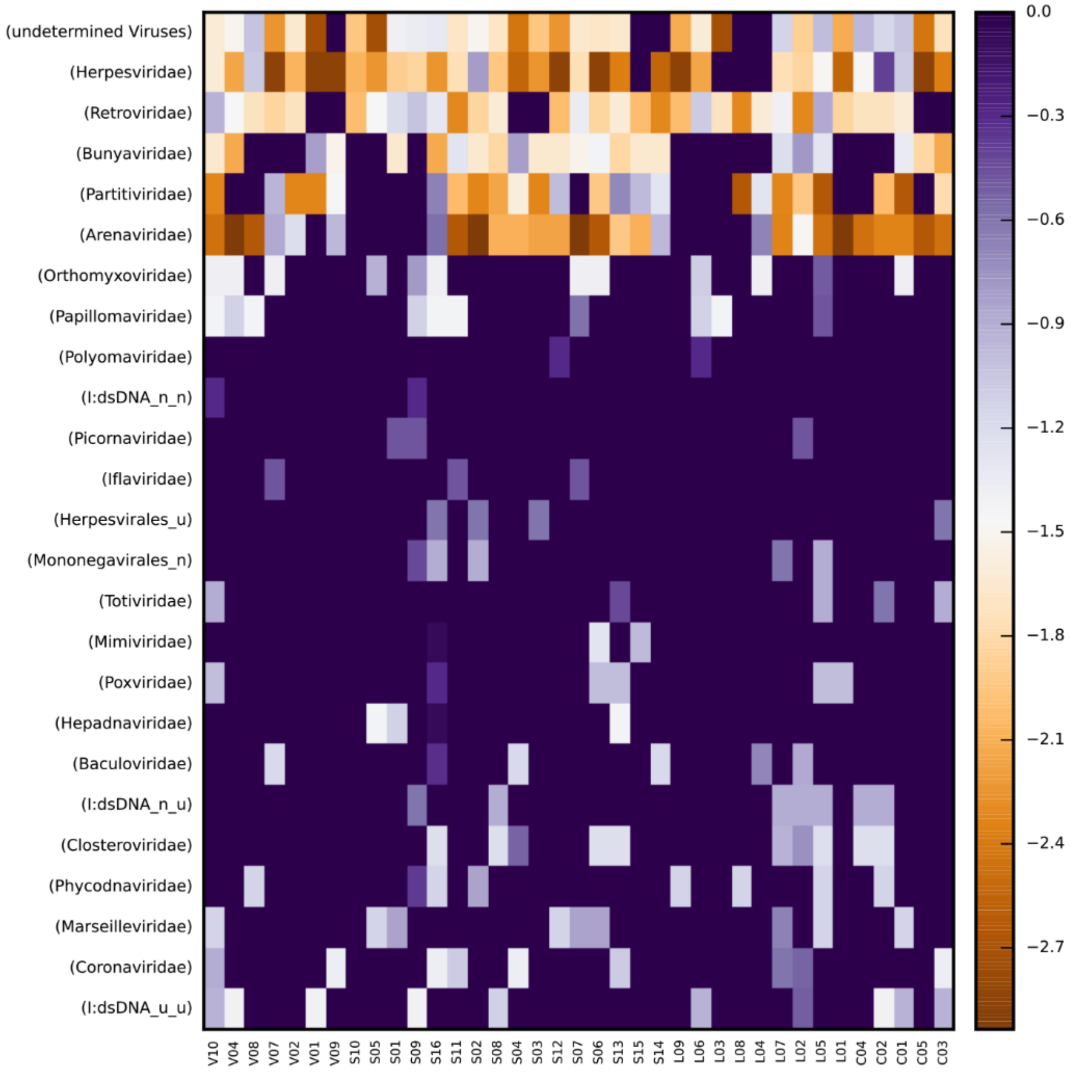


Figure 74. cDNA viral subset – Heatmap at the family level.

The DNA viral subset was also examined to detect markers contributing to differences between pathologic groups. The Baltimore classification (Figure 75) showed a lower amount of ssDNA viruses but was not a unique feature in a single group. Sequences with no classification coming from phages were also widespread among all samples. Other tables at different taxonomic levels are found in the online repository.

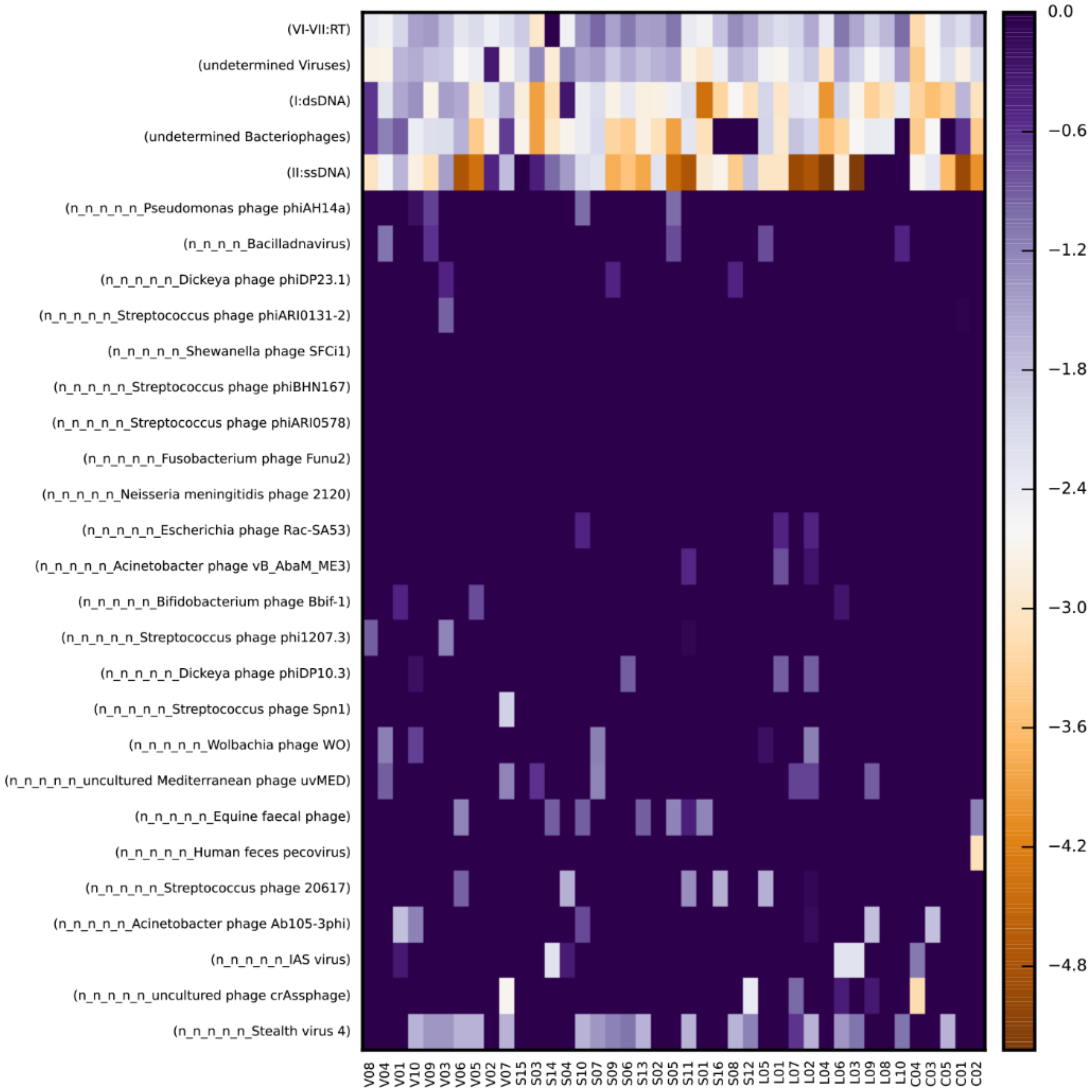


Figure 75. DNA viral subset – Heatmap using the Baltimore classification.

At the family taxonomic level (Figure 76), the OSCC and OL had a lower prevalence of some of the main bacteriophage families *Caudovirales*, *Siphoviridae* and *Myoviridae* which were present more prominently in controls.

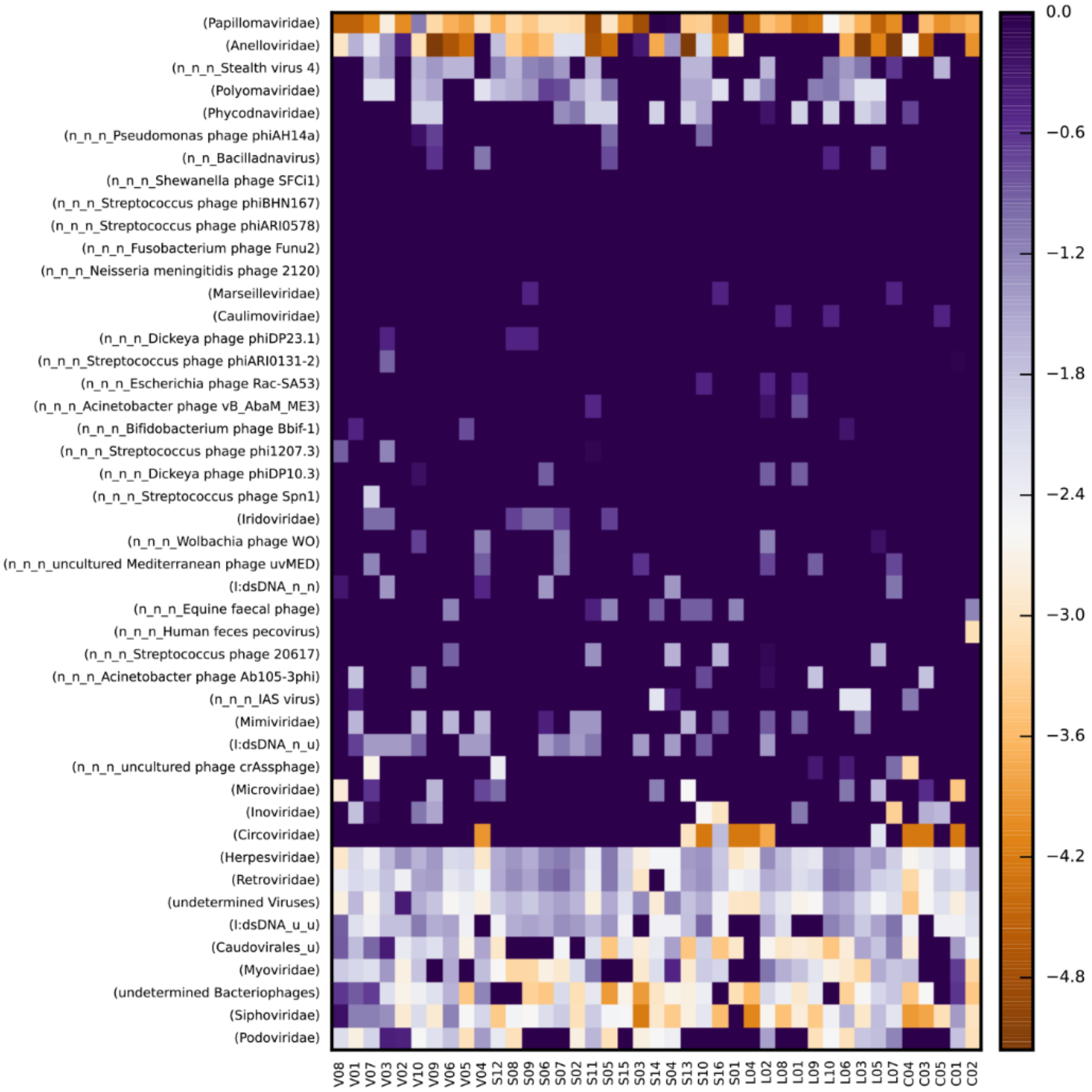


Figure 76. DNA viral subset – Heatmap at the family level.

For the 16S dataset (for which the Jaccard matrix supported cluster formation), at the phylum level, the OL and OSCC groups had a lower prevalence of Synergistetes and Tenericutes. Spirochaetes and TM7 were also differentially distributed among samples (Figure 77).

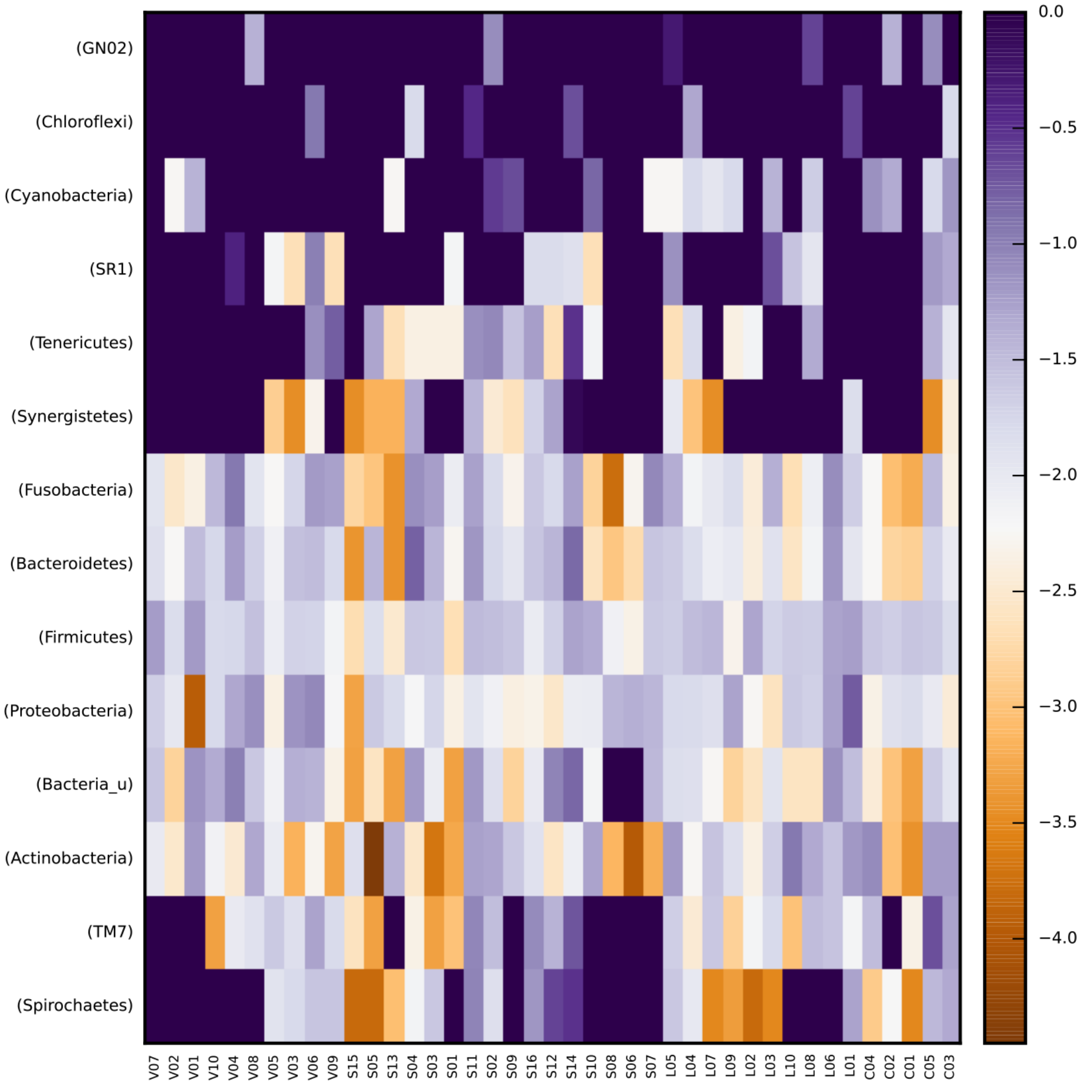


Figure 77. 16S Dataset – Heatmap at the phylum level

The cDNA non-vir subset heatmap at the phylum level (Figure 78) displayed differential abundance in Firmicutes, Bacteroides and Fusobacteria with less prevalence in controls and OL samples, as well as hits from some fungi such as Basidiomycota and Ascomycota. Verrucomicrobia were more prevalent in non-controls.

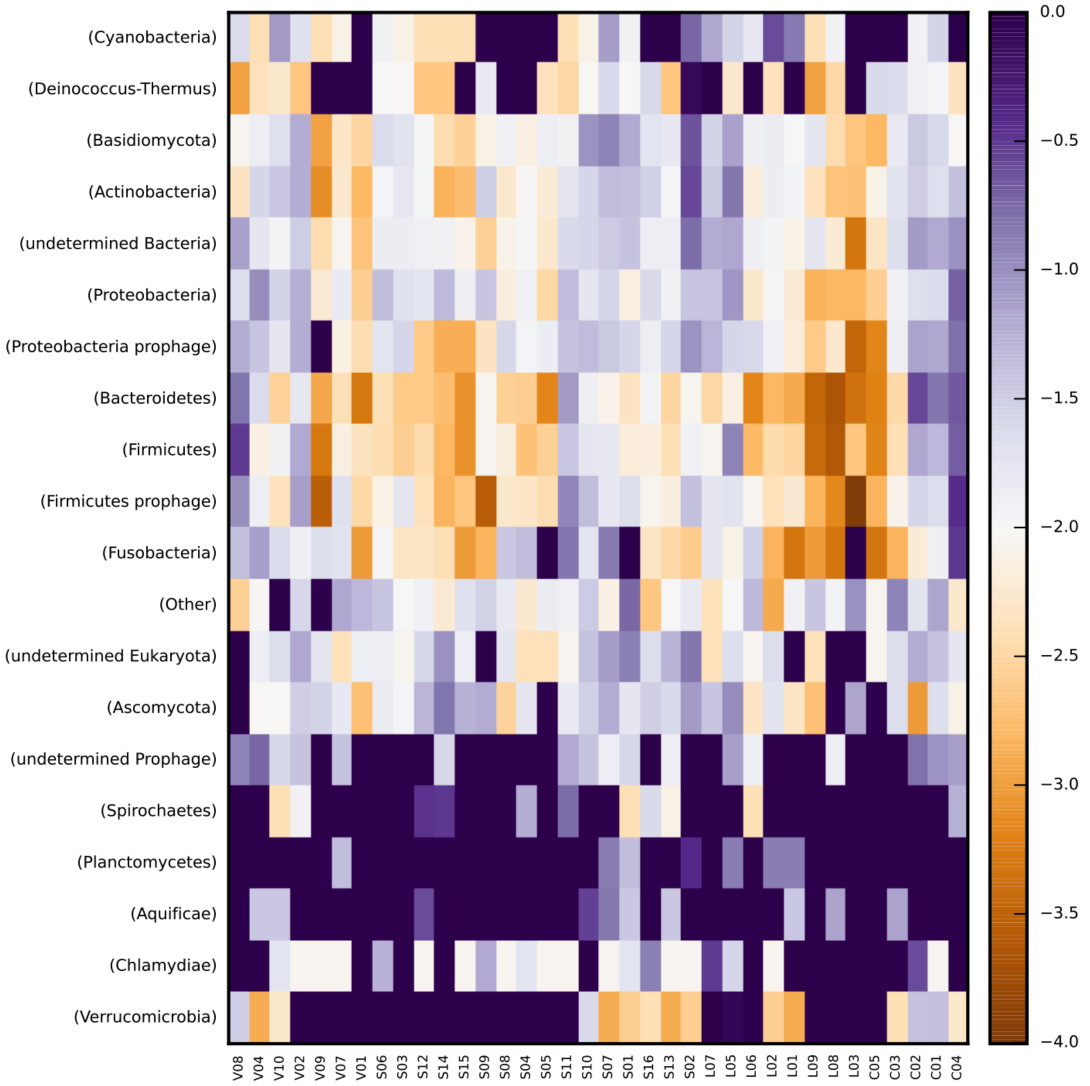


Figure 78. cDNA non-vir subset – Heatmap at the phylum level.

Finally, the DNA non-vir subset at the phylum level (Figure 79) presented a differential abundance of *Ascomycota* and Proteobacteria prophages (specially in OSCC samples). Verrucomicrobia was also found in lower proportions in some of the PVL samples.

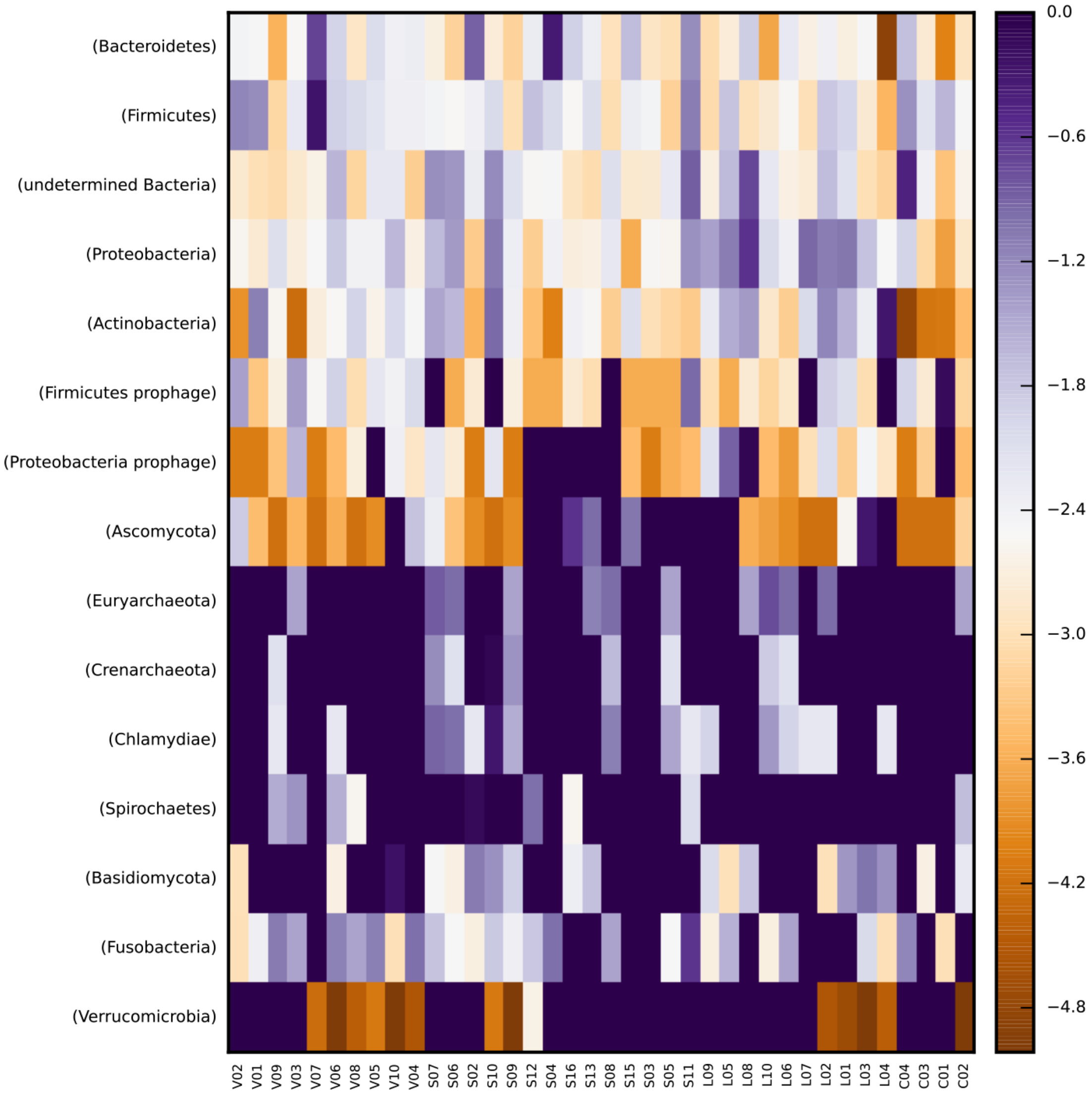


Figure 79. DNA non-vir subset – Heatmap at the phylum level.

The complete collection of heatmaps produced for each taxonomic level in every dataset can be found in the online repository https://github.com/rodrigogarlop/PVL/ in the Heatmap folder.

Evaluation of differential abundance. Heatmaps are useful tools for evaluating records that are present in different abundances in the datasets but more powerful automated methods exist for evaluating differential abundances of groups and categories within groups. In order to evaluate the differential abundance organisms and viruses in the current study, three approaches were considered.

The first one was to analyse our tables using a linear discriminant analysis (LDA) using LEfSe. In general terms, LDA works similarly to the linear models used for dimensional reduction in the beta diversity. They try to model a set of differentially abundant markers (more or less prevalent in a group) to evaluate which ones contributes the most to the observed differences (see Methods for some more details). The resulting LEfSe produced a list of differentially abundant taxa (Figure 80) that were overrepresented in each of the different groups. Unfortunately, the viral tables could not be used with this approach as it seems there were too few differentially distributed items in both tables, yielding void results.

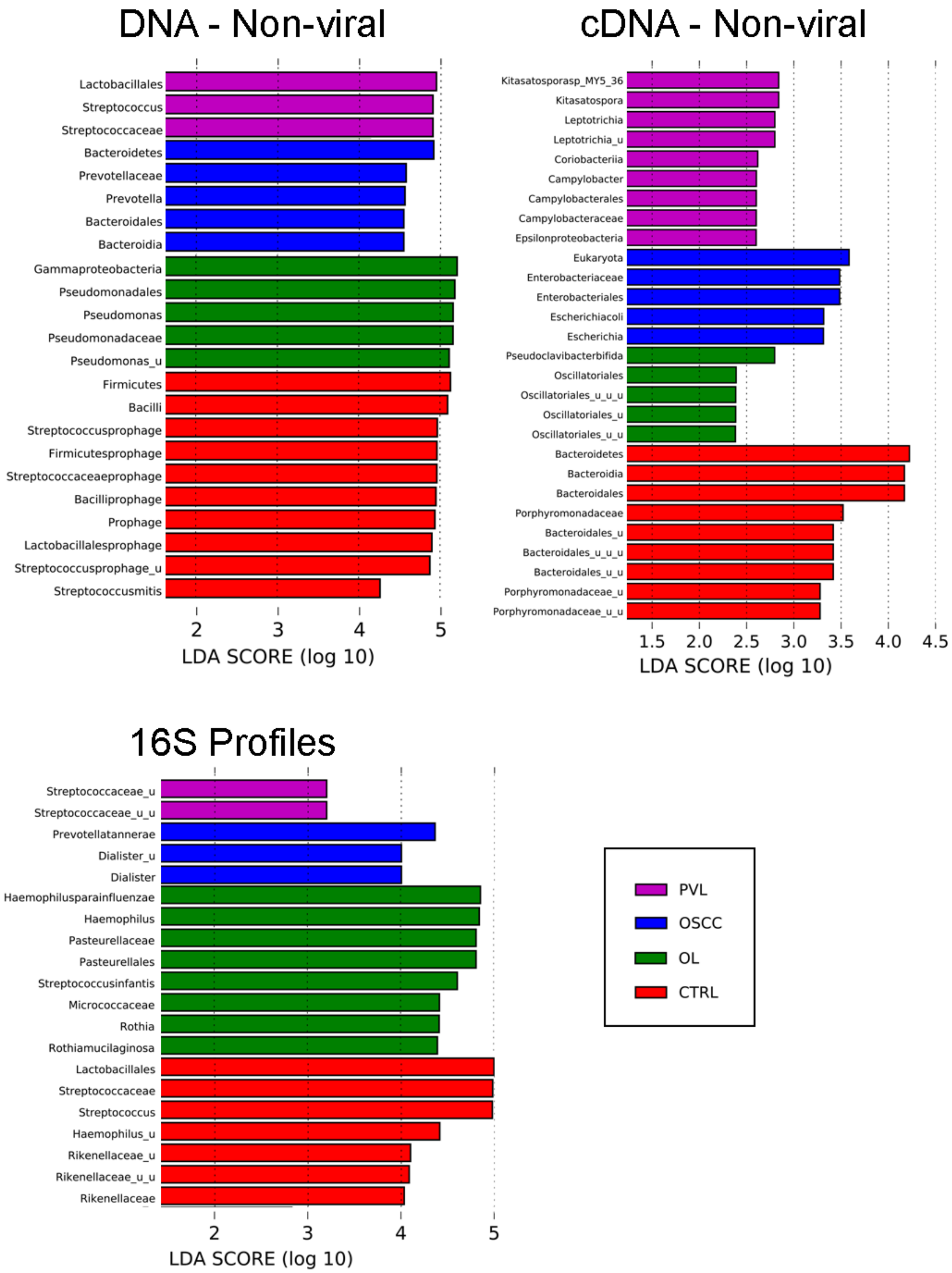


Figure 80. linear discriminant analysis (LDA) using LEfSe.

In the 16S profiles, unidentified streptococci were the only markers that were detected as differentially abundant in the PVL group. For the OSCC group, the *Prevotella tannerae* family and *Dialister*, for the OL group the *Haemophilus parainfluenzae*, *Pauseurellacea*, *Rothia* and related taxonomic ranks. In control group, Lactobacillales and streptococci. The two were precisely the ones that were identified for PVL in the DNA non-viral set, Bacteroidetes in OSCC and Gammaproteobacteria, and *Pseudomonas*-related ranks in the OL group. Finally, Firmicutes and prophages in the Control group.

The cDNA non-viral group presented lower scores, with *Kitasatospora* in PVL, Eukaryotes (fungi) in the OSCC and *Pseudoclavibacter bifida* in OL. Finally, Bacteroidetes in the Ctrl group.

Other two methods were employed for assessing viral differential abundances. The first was a Kruskall-Wallis test to compare items in the tables based on group separation but gave no results for viruses; only for bacteria (tables available in the repositories).

Finally, the DESeq2 method for differential abundance was used for identifying such markers in the viral sets (as no LEfSe results were available). The method gets a list of differentially present items in paired comparisons. Thus, all permutations in the group category were used and results were filtered with an 0.05 pvalue. The only viral markers that were found in the cDNA viral subset were in the comparison of OSCC and control: *Pittosporum cryptic virus*-1 and the undetermined viral abundances (which has no relevant taxonomic information). In the DNA viral subset, several viruses were found to be differentially abundant: in the OL versus control, several streptococci, *Ralstonia* phage, an Eel pequenovirus and several with undetermined categories.

The comparison of OSCC and control for the same set also throwed several of the same viruses, plus *Torque teno virus* and *Human papillomavirus*. In the OSCC versus OL there were differential Torque teno viruses. In PVL versus control the same *Streptococcus* phages, *Torque teno virus* and *Human papillomavirus*. From the PVL vs OL some more *Anelloviridae* with no label, as well as various phages and endogenous virus (that must be from transposons). And finally, from the PVL versus OSCC was populated with significantly abundant phages, retroviruses and *Human endogenous* viruses. The same comparisons were carried out for bacteria and produced lengthy lists of differential abundance biomarkers.

The complete collection of LEfSe results produced for all non-virus datasets and the tables comparing differential abundance can be found in the online repository https://github.com/rodrigogarlop/PVL/ in the Differential abundances folder.

DISCUSSION

Since it was first described in 1985 by Hansen and collaborators, the proliferative verrucous leukoplakia has remained a puzzling challenge for both clinicians and researches alike (*207*). As of today, no aetiology has been described and its malignant nature makes it a relevant topic to address.

As other malignant lesions, particularly in the oral cavity, the understanding of the onset and progression of the lesion would provide advantageous as diagnostic is complicated and is often achieved only after the lesion evolves, frequently linked to the development of OSCC (*201*, *213*). Thus, the importance of the search of an aetiological agent which this study has focused on.

Previously, researches have tried to look for a viral agent with mixed results ranging from the discovery of oncogenic viruses reported in some studies (*223*) to larger surveys yielding no clear aetiological agents (*224*, *225*), as was the case of the screening we carried out prior to the development of the metagenomic analysis.

This is mainly explained by two reasons: first, the parallelism between the cervical cancer and oral cancers (*216*, *223*) and the recurrent and multifocal nature of other viruses involved in similar lesions (*222*, *340*), most importantly, HPV, which was one of the first to be surveyed in relation to PVL in a 1995 paper by Palefsky and collaborators (*223*). The search for the aetiological agent has also included other oncoviruses such as the EBV(*227*, *341*).

The first study of this project, the PCR screening, was designed as a thorough exploration of the potential oncogenic viruses that may be found in the oral cavity. Some of them had been reported in the oral cavity but most were just selected for their potential role in this type of cancer in particular. Although screenings had been carried before, they were limited to specific viruses.

Instead, our study aimed at using available “universal” degenerate primers for means of a broader selection of papillomaviruses, polyomaviruses and herpesviruses species in patients presenting various diagnostics, an additional improvement over previous works. OSCC and OL samples were included to compare PVL with related afflictions.

More specifically, OL biopsies were included to contrast a condition producing related similar symptoms and OSCC was used as a way to study the progression and has been previously been reported to be affected by viral infection (*342*), as most PVL lesions develop into these type of cancer over time (*206*, *207*). Additionally, the viral survey was designed as a blind study to prevent methodological biases in the laboratory procedures and the bioinformatic processing.

The result, a negative one for all primers employed, as none of the targeted viruses was detected in any of the samples, was in accordance to other previous studies that used specific primers targeted at HPV (e.g. type 16) (*217*, *224*, *241*). Although the HPV variants have been proposed as influencing the development of OSCC, this observation has not been confirmed (*343*). Yet, this was expected and was a stepping stone for the larger metagenomic survey that was carried out afterwards and served as further justification of the need for a wider type of analysis that may get a broader panorama of the prevalent viruses in the samples.

Limitations of the study include the fact that no laser capture microdissection was employed during the clinical extraction of the biopsy and therefore it cannot be discarded that some of the extracted DNA may instead be from tumour infiltrating inflammatory cells or from stroma. This limitation also applies for the second study, the metagenomic exploration.

A major caveat was that no positive control for polyomaviruses was available at the time of the study. To compensate for this, more bands were excised and cut as a precaution, some of them in duplicate for sequencing. Plus, polyomavirus amplicons were also generated *in silico* with the Primer Prospector software using the available genomes to predict the fragment that would be produced.

Homology searches are heavily dependent on the database that is employed and the search algorithm, with the BLAST suite still posing in the scientific literature as the favourite approach to carry them out. Newer approaches have made large database feasible but blast's precision was adequate for the scope of the analysis.

Also, depending on the algorithm that is employed, results may vary. Megablast may be used for the initial searches as it is faster and works well with very similar sequences, which were expected to be recovered from a family database. For searching more distant homology, tblastx is more suitable. In this type of blast, searches are carried out with translated nucleotide queries against translated nucleotide databases. As a result, nucleotide sequence conservation becomes secondary and distant hits may be found.

The downside of this algorithm is that it is computer and memory-intensive, particularly for searches against large databases such as the complete NCBI's nonredundant (nr) which are some of the reasons why the blastn algorithm, a regular nucleotide-on-nucleotide algorithm was employed for the most general screening against all organisms. When compared to megablast, blastn sacrifices velocity for an increased sensibility and, in general, finds more results which are still higher in specificity than tblastx's, which was the second reason why this was used instead as we wanted to avoid overly distant hits.

The reason behind using the family databases resides in broadening the search to include related species whereas the nr inclusion means a last resource for sequencing not matching know databases. An observation that applies to both studies in the project is that the resulting identifications are just as good as their reference datasets, meaning that larger and more diverse databases provide an ampler spectrum of the possible sequences that may be present. This, however is often one of the main limitations in viral metagenomics since most viruses rely on incomplete, or generally less than optimal sequence databases for identification of viruses.

By-products from the PCR in the screening, mostly identified as human contamination, are expected to occur, given that the first step of the process consists on tissue disruption. The same is true for the second study. Samples with hits against humans are thus not entirely unexpected as a result since biopsies are the source material for DNA extraction.

In the case of the screening study, extraction was made with phenol-chloroform, one general purpose extraction protocol along with proteinase K that is has been reportedly used for obtaining nucleic acids from viral particles by multiple research groups (*48, 246*). During extraction, however, the processed tissue that remains after the homogenization step by mechanical tissue disruption behaves very differently during the first steps of the extraction, even after the pellet removal. Despite correct sample processing, it also introduces bias that may impact results (*344*). Consequently, as an improvement for standardizing the extraction process, RNA and DNA in the second study were processed with the QIAamp viral RNA commercial extraction kit, effectively reducing the experimental bias. Although it has been published that it may produce a smaller total yield (*345*), the extraction protocol ensures a correct extraction, as was confirmed in our study.

Sample hardness is also a factor that may impact extraction negatively or at least introduce a bias due to the sample source. The bead beating protocol is an effective way to disrupt tissue in general but had a different rate of success, mostly depending of the hardness of the tissue. As most hyperkeratotic patches are much harder to homogenize they required more time in the Tissue Lysser, which in some cases leads to tubes breaking (they were covered in parafilm to prevent leaking). The resulting homogenization may thus vary in success, with some biopsies (specially the small and hard ones) remaining seemingly unaltered after the process. Alternatives to this process such as TRIzol extraction would require proteolytic enzymes but may introduce other types of biases such as coextraction of plasmid DNA and RNA (*346*).

It is important to note that all the experimental protocols were originally designed with 454 pyrosequencing in mind and were adjusted accordingly as the Illumina platform became widely adopted following the 454's discontinuation announcement (*35*). Several steps in the protocols had to be adjusted accordingly. The most important one was the total DNA concentration in the libraries but also features related to the sequencing construct structure (total length, primer selection, library construction, etc.) and other procedures that were carried out originally for pyrosequencing. Some could be adapted to Illumina but bearing the cost of a series of issues in the downstream process.

The construction of libraries for 454 sequencing requires a large input DNA concentration, which, for the original protocols was set at around 200-500 ng of pure DNA, although concentration-efficient methods have been published for using a lower amount and still get a decent amount (*347*). On the other hand, the Illumina library construction requires ~5-50 ng of pure DNA for library construction depending on the type of library. Since the study was designed originally for the 454 methodology, the total DNA was amplified using the SISPA and GenomiPhi approaches for the RNA and cDNA fractions respectively, producing far more molecules than ultimately required as both of them were sequenced in an Illumina MiSeq platform.

The 454-FLX with Titanium reagents allows for the sequencing of 500-700 bp fragments although, in practice, they are more frequently on the inferior bound (the actual mean of the amplicon set was in fact less than 500 bp). Considering this, for the 16S amplicon libraries, the fragment was selected to be amplified with primers E8F and B530R, which amplifies a sequence of ~550-600 bp which, along with the added MIDs produces fragments that are not adequate in size for Illumina sequencing (no overlap would have been observed for most sequences in Illumina TruSeq libraries and they were not long enough for the introduction of indexes by transposase in Nextera libraries).

With this in mind, the amplification for the DNA set was carried out with GenomiPhi, which has been previously reported to amplify circular ssDNA genomes preferentially (*348*) whereas for the cDNA, a SISPA amplification was carried out, producing smaller fragments that were then problematic for the construction of the Illumina libraries. A recent work by Kugelman and collaborators exploring different methods for viral RNA preparation (*349*) has issued a warning regarding SISPA's preparation error rate. In fact, the current sequences represent our second attempt at sequencing the cDNA dataset because we first used a Nextera XT approach with no viable results (due to the short fragment length, the rate of success of the index introduction is decreased significantly resulting in even shorter inserts that are not adequate for sequencing).

Bioinformatically, the greatest challenge consists in filtering the indexes that are introduced by the SISPA random primers. The hexamers produce amplification but also small tandem repeat constructs that have to be filtered bioinformatically. Also, the 20 nt sequences in each end imply the removal of 40 nt from each of the sequences in order to get the actual usable sequence.

Due to the short size of the resulting sequencing reads, the cDNA may require further bioinformatical processing to recover sequences that cannot not be identified normally. This is possible because the construct of short sequences was completely sequenced in the 200 cycles that were used (if the insert size is smaller than the total cycles, the 5'-end of the construct, the sequence flanking that includes the index, gets sequenced). The downside is the small size of the insert and that, whenever the amount of short sequences is large in a whole library, there is a high risk of the Illumina apparatus aborting the run previous to the sequencing of the first index. However, we managed to recover a considerable number of usable reads that would have been discarded otherwise.

We also have developed different databases and bioinformatic search strategies for tackling some of the major problems with viral metagenomics: the incomplete sequence databases (*260*). There is a marked bias towards sequences from commercially or clinically relevant pathogens in databases such as ViPR (*350*), some of them contained in specialized curated databases such as the OpenFluDB (*351*), and the Los Alamos HIV sequence compendium (*352*). Although important efforts have been made towards a more standardized knowledge like the ViralZone (*353*) taxonomically-curated resource and comprehensive new databases integrating metadata such as viruSITE (*354*) database, most of these have not been created for large-scale explorations as in metagenomics.

Despite there being repositories in the International Nucleotide Sequence Database Consortium (*355*), the main organism coordinating sequences form three main repositories (European Bioinformatics Institute, the National Center for Biotechnology Information and the DNA Data Bank of Japan), for every viral sequence that is reported, many of them have no relevant taxonomic information as they are part of screening analyses and just have partial sequence available. Our database construction approach aimed at compiling the largest number of viral nucleotide sequences available and clustering the set with reference genomes to reduce redundancy while favouring variability by retaining cluster-independent sequences.

Due to the small number of sequences in some samples and the short size of part of the usable sequences, methods had to be developed to overcome this limitation. Our selected approach relies on assigning taxonomy directly to the reads, with no prior assembly but using an LCA approach, similar to the ones used for taxonomic classification in programs such as MEGAN (*356*) and more recently Kaiju (*357*), but using different databases spanning other domains of life, which leads to the survey of the fungal and bacterial fractions. The sequences searches can be executed equally for the databases using at least 80% of the sequences in alignments so that filters would be strict. Only results that have no hits across databases are utilisable in the final datasets. This approach allows us to keep only uniquely identified viral, bacterial, archaeal or fungal sequences, after human filtering. However, archaeal hits were not sufficient to remain for the statistical analysis in the study.

Taxonomy is also an important issue because the existing viral sequences have several missing ranks, due to their nature and the old classification systems which have remained in use (*259*, *260*). Most importantly, the inclusion of the Baltimore classification for most records in the database allows for a further classification of the types of viruses according to their type of nucleic acid and replication mechanism. The usage of a special nomenclature system with missing ranks being relabelled also allows avoiding them from joining unranked pools containing several different viruses.

The homology searches and filters were carried out with conservative parameters to avoid spurious identification. The downside of this approach is that the total number of non-identified sequences may increase (lower sensitivity) but a higher specificity can be achieved.

Perhaps the greatest challenge in the viral analysis is working with the contingency tables as the low totals in the sets after filtering require careful processing to avoid biasing result due to differential sequencing depth. Yet, we managed to study not only the viral fraction but the fungal and bacterial as well from the same datasets. As a result, we have been able to get a broader picture of the actual microbiota, which is in fact a very complex community spanning al domains in the tree of life and beyond.

Our analysis of the bacterial distribution at the phylum level, based on the 16S profiles, has shown a higher prevalence of Firmicutes in all samples, as well as of Actinobacteria, Bacteroidetes, Fusobacteria and Proteobacteria. This is in accordance to most studies of the oral microbiota (*116*, *169*, *173*, *191*), and is supported by both cDNA and DNA WGS sets, although Firmicutes rank in different positions as they have a higher prevalence of Actinobacteria and Proteobacteria, respectively. The cDNA set is particularly interesting as most sequences come from rDNA, meaning they partially reflect the actual active fraction of the domain (although no differentiation was made for other types of sequences from different origin).

At higher taxonomic levels, the abundance is not reflected across the three bacterial datasets, with a higher abundance of *Pseudomonas* and *Propionibacterioum* in the WGS sets. The former is regularly detected in saliva and is a member of the Proteobacteria phylum commonly associated with chronic infection of the respiratory tract. The latter is a normal commensal and opportunistic pathogen in endodontic infections (*358*). The 16S analysis, on the contrary, shows a higher prevalence of *Veillonella*, a lactate fermenting streptococcus that is a common commensal in the oral mucosa and biofilms of the tongue and plaque due to an active role in coaggregation (*359*). The *Rothia* genus, a normal benign commensal that is rarely implicated in endocarditis (*360*) is also in higher prevalence in the 16S set, along with *Leptotrichia* and *Haemophilus*, which are also reported in the HMP datasets as regular commensals (192).

Our results have shown that the most prevalent viruses differ markedly in both WGS datasets. The cDNA set presented significantly higher counts of *Mammarenavirus*, a rodent virus (*361*) that has been reported to be human transmitted to humans (*361*). Regardless, as it happens with all database searches, it is important to note that the sequences found may be just similar to genus and not the exact same type. The same consideration must be taken into account when analysing any results from metagenomic studies. Another family showing high prevalence, the *Partitiviridae* family comprises by non-enveloped viruses normally infecting fungi that may be related to the actual fungal population (*362*).

The fungal microbiota (or mycobiome) is also poorly understood as it has been ignored in most studies. However, it has recently received some attention such as in a study by Ghannoum and collaborators of the oral healthy mycobiome, one of the largest so far (*363*), although focussing on specific internal transcribed spacers as markers. This study is, to our knowledge, the first attempt at studying the oral metagenome of fungi associated to the OSCC, OL and PVL lesions. The surveyed fungal population is representative of the reported oral fungal community, including *Candida* species, *Cryptococcus* and *Malassezia* in both Illumina datasets (*364*). In the DNA set, the *Yarrowia* , a poorly characterized genus was also detected.

Previous specific studies have failed to detect a connection of the PVL development and infection with species of *Candida* (*216*). It is important to note that sequences that are aligned to the references of Ascomycota and Basidiomycota, from superkingdom Dikarya (of “higher” fungi), were detected but the LCA pointed to their phyla only. This is congruent to recent analyses uncovering the healthy human mycobiome (*365*). The fungal population is estimated to be at around 700 different taxa in the oral cavity (*365*).

In the case of the DNA viruses, bacteriophages are widely spread as it would be expected for predation of local bacteria. Congruently, the most prevalent families were those in *Caudovirales*, *Siphoviridae* and *Myoviridae* which is congruent to previously reported phages in the oral cavity (*366*). In a finer detail, the local phage populations, corresponding mostly to streptococci phages would be congruent to the high population of their target bacteria, which are also some of the most abundant bacteria in all the groups (OSCC, PVL, OL and Control) as supported by the LEfSE analysis.

Also from this analysis, controls exhibit a higher abundance of Bacterioidetes, namely of the *Porphyromonas* genus Although no species level resolution is obtained for most taxa, we detected a species related to *Porphyromonas gingivalis*, a pathogen associated to periodontitis that may also be found as part of oral microbiota (*367*). It is important to stress that the observation of specific species does not necessarily imply these exact agents may be present in the samples as homology search can only provide information on how similar a sequence is to a database of references. The closer a sequence is to such reference, the better the score the hit will receive but homology cannot be used as a means of undoubtedly pointing out to the existence of a particular species.

Although beta diversity in our analyses shows no statistically distinct sample clusters in the viral populations that could be group related, there is a single statistically significant abundance difference in the paired (not by group but by dual categories) comparison of OSCC and the control group showing a higher prevalence of a virus from the *Partitiviridae* family, a *Pittosporum cryptic virus* 1. Such family is composed of virus infecting fungi and plants (*368*), which may suggest it may be related to a different host, possibly fungi. Its role in a human associated environment, however has not been described.

Other significant results in the DNA viral set include, for instance, the OL samples having a statistically significantly lower abundance several types of bacteriophages, including *Streptococcus phage* PH10, *Streptococcus phage SM*1, and *Streptococcus phage EJ*-1, the same three that were statistically supported in lower abundances in the OSCC with respect to the control group. This would support the idea of the viruses playing a significant role in the control of bacterial populations, supporting a potential relevant ecological role in terms of the overall population of bacteria that has been hypothesised before (*368*).

Precisely, when analysed for differential abundance, the most relevant LEfSe markers in the PVL group was that of streptococcus. No statistically significant difference was detected, however with the matching bacterial population. That is not for the same grouping at least because streptococci are widespread among all samples, suggesting a moderate impact in the resulting bacterial population. In the 16S profiles group, genus *Dialister* and *Prevotella annaerea* were differentially more abundant in OSCC (LDA score ≥ 4).

DNA viruses statistically more prevalent in the different categories of the groups include an unclassified HPV and HTT in the OSCC relative to the control. It is important to recall that HTT is a virus that has a wide prevalence in the whole mammalian population and is was reported to be found in blood (*247*, *248*). However, due to the use of the MDA amplification, it is documented to be preferentially amplified along with other small sized circular ssDNA viruses (the genome of HTT is 3.8 kb long) (*369*, *370*). In fact, in a 2014 paper, Erick Wommack's group go as far as urging the metagenomics community to avoid MDA techniques in favour of an unbiased sequencing (*348*), something feasible under the current Illumina sequencing paradigm but unreasonable in the 454 era and during the initial planning of this study.

In spite of detecting no statistically supported differences between the groups (FDR adjusted p-val > 0.05) formed in the alpha beta diversity analyses, interesting oncogenic viruses were searched within the sample composition (contingency tables) of the DNA viral subset. Interestingly, *Human Alphapapillomaviridae* 9 (which includes the oncogenic types 16 and high-risk types 31, 35, 52, and 58) is found with no prevalence for specific groups among the viral species of 36 of the 41 samples. Its total count was a mere 674 hits but distributed across all groups with no specific prevalence in neither. This observation does not support the estimates of 1% prevalence in large cohorts that has been previously reported (*371*) but the metagenomic approaches do not have the specificity to determine species in most cases, let alone the type. No other high-risk variants have been reported to be present in the samples but many could not be assigned beyond the family level. Still, papillomaviruses are extensively distributed through all samples. In total, eight different types or species of *Papillomaviridae* (many undefined at species level) were detected in at least two samples with over 200 total counts in total.

Also interesting but not significantly different for a particular group, the prevalence of HHV-4 or EBV as the DNA contingency tables report them to be present in the 33 of 41 samples regardless of the group. This observations supports similar findings by Bagan and collaborators (*341*). Nonetheless, the low observed absolute values were evenly distributed across our groups.

Recent discoveries have associated temporal infections by viruses to the development of immune disorders, such as the discovery of the long-term effect of infection by reoviruses in the development of celiac disease, reported by Bouziat and collaborators (*372*). They suggested that acute infections may be able to trigger a pathology in the future. Although different in nature, it cannot be ruled out that a similar mechanism may be operating in the onset of PVL. In this regard, potential causal agents may be not identified because, at the moment the samples were obtained, they may have decreased in number to become undetectable or remain concealed among other constituents of the virome. However, its effect may eventually become patent. Because of this temporary aspect, a follow-up study with different time points may prove very valuable as it may shed a light on the actual development of the disease, complementary to the static snapshot that this retrospective study represents. Nevertheless, resolution to certain constraints must be considered for this approach. These include, but are not restricted to the complexity of sample collection, mainly due to the invasive nature of biopsy removal compared to other types of samples (stool, saliva, etc.), as well as the lack of a genetic marker associated to the PVL that may be used for narrowing the search in the general population for the selection of a cohort of potential candidates for this longitudinal prospective study, before any lesion-related features from the major criteria for diagnosis can be detected. Finally, the complex PVL diagnostic may be a critical limitation but the changes in the microbiome occurring at the onset of OSCC may be assayed, as well as the progression associated to the appearance of new leukoplakia patches after the removal of the affected area (lesion recidivism).

Even if no relevant associations are found for the groups and the underlying microbial and viral communities, the analysis has shed a light into the ecological spectrum spanning across different domains and statistically support markers may be further analysed in further studies. However, it would be important to emit several recommendations regarding the experimental procedures that may be employed for viral analysis.

The most significant challenge in the bioinformatic study is the initial total number of usable sequences. This is, in great part due to the large proportion of human and bacterial sequences present in all the samples, which is mostly experimentally related. Further analyses must be designed with Illumina analyses in mind to avoid amplification biases of the SISPA and MDA. The database search may be updated and improved with newer available sequences from the recent IMG/VR database from the Joint Genome Institute (*373*), currently the largest publicly available database including sequences directly derived from exploration of viral sequences within metagenomes.

A large proportion of the reads remains unidentified as taxonomy assignment is mainly limited to the sequences comprising the collection of references. The percentage of this fraction is highly variable between samples and is particularly prevalent in the cDNA WGS set (over 60%) or less stringent parameters. These sequences may be further explored with different databases as they may still contain relevant markers differentially distributed among the groups in the study. As new species become sequenced and characterised, those unidentified reads that account for significant proportions in most of the current datasets may be eventually unravelled. Therefore, the importance of this "dark matter", though undetermined, cannot be discarded. This fraction may be worth exploring since it still may contain interesting markers.

Also for future analyses, exploring CRISPR spacer sequences and tRNA from viral origin in the bacterial fraction may throw interesting data regarding the relationship of bacteriophages and their prey counterparts as they pose as viral signature of the species that may have caused infection in their evolutionary past. These convenient structures have been exploited for viral-bacterial association discovery by Nikos C. Kyrpides' group (*293*). Hidden Markov Models published by the same group may also provide a more flexible approach than regular homology search by sequence alignments as they consider transition states of the sequences.

The WGS approach employed for the sequencing of both DNA and cDNA datasets enable the exploration of functional profiles using sequence assemblies. These are currently being carried out as part of the study in the hopes of unveiling relevant information regarding the functional components that may differ between the OL, OSCC, PVL, and control groups. These results have not been included in the manuscript due to time constraints and technical issues, although they will be submitted for publication in the near future.

CONCLUSIONS

- In order to detect an aetiological agent for the oral precancerous lesion known as proliferative verrucous leukoplakia two studies were carried out.
- A broad screening study was completed for assessing the presence of oncogenic viruses in 40 samples with four sets of oral biopsies from patients diagnosed with oral leukoplakia, oral squamous cell carcinoma, and proliferative verrucous leukoplakia using different collections of broad-spectrum primers to amplify potential fragments of oncogenic viruses in the samples. In this blind study, no amplifications yielded any detectable sequences for Human papillomaviruses (including oncogenic *Human papillomavirus* types 16 and 18), herpesviruses (including oncogenic *Epstein Bar virus* and *Kaposi's sarcoma-associated herpesvirus*) nor polyomaviruses (including oncogenic *Merkel cell polyomavirus*).
- The existence of a potential viral agent for the lesion cannot be discarded. This study served as the stepping stone to carry out a broader taxonomic survey of the metagenomic content using a new set of patients in a larger metagenomic and 16S profile study that was carried out to study the viral and microbial fraction of the oral cavity.
- The second study was carried out using 41 new samples from the same four types of lesions. DNA and RNA were extracted for 16S profiling of the bacterial fraction, which was carried out with 454 pyrosequencing using primers directed to the V1-V3 hypervariable regions. The DNA fraction was enriched with a multiple displacement amplification protocol and sequenced for 2x300 paired-end reads in an Illumina MiSeq platform. The RNA fraction was enriched for the nonspecific amplification of viral sequences using a MDA and SISPA protocols, respectively, and sequenced in an Illumina MiSeq platform for 2x200 paired-end reads.
- New custom-made databases were created for the analysis of the viral, bacterial, fungal and archaeal fractions, along with their taxonomic references, and bioinformatic pipelines and scripts necessary to process the taxonomic assignment. The whole genome sequencing datasets were used to explore each of the domains and the viral spectrum, managing to unravel part of the “dark matter” in the samples.
- The taxonomy of all datasets was standardized to fill the empty ranks with special strings to trace the nearest valid taxonomic label so that even unclassified pools could be differentiated.
- The viral dataset was specifically tailored to include useful non-ranked taxonomic categories such as the Baltimore classification.
- The SISPA and MDA methodologies encompass a series of disadvantages (e.g. in the former, shorter usable reads, production of chimeric sequences, and formation of artificial tandem-repeated library constructs; in the latter, a bias towards the amplification of ssDNA and circular genomes and the production of short sequences) that discourage the use of such techniques, favouring amplification-free approaches as better alternative.
- Globally, in the 16S profiles set, the most abundant bacteria belong to the phyla Firmicutes, Bacteroides, Fusobacteria, Proteobacteria, and Actinobacteria. However, in the subsets from the WGS collections, the abundance of these phyla ranked differently, with Actinobacteria, and Proteobacteria as the most prevalent.

- Sequences in the cDNA viral subset had a higher number of matches with bacteriophages than eukaryotic viruses, which was also observed for the four groups in the study. The DNA viral subset had a higher prevalence of eukaryotic viruses globally.
- Inter-individual variability of the species/OTUs abundance in all datasets was higher than that reported between the different groups, contributing to a diminishing resolution capability of the statistical analyses between groups and the differential abundance comparison between the records.
- In both WGS viral subsets, the low number of reads that were successfully assigned a taxonomy was markedly uneven among samples, preventing the usage of regular scaling for data normalization and forcing the sets to be rarefied for even comparisons. The non-viral subsets, as well as the 16S set, did not present this problem and were transformed using cumulative sum scaling instead.
- This study contains the first metagenomic analysis of the fungal fraction associated to these pathologies.
- The alpha diversity analysis showed a bacterial distribution that is congruent to oral taxonomic studies but the separation of the groups was not statistically supported for the composition of the pathology groups in the study as shown in the beta diversity analysis.
- The different fractions were analysed in search of groups of markers that reflected separation by groups with no statistically relevant groupings.
- Several methods were tested to identify statistically relevant differential viral markers in the PVL and OSCC groups but ultimately resulted in no etiological relevant marker in neither.
- Oncoviruses were surveyed in the contingency tables of the DNA fraction, identifying HHV-4 and high-risk types of HPV in most samples but no relation to the group distribution could be statistically supported.
- No aetiological agent could be determined by the metagenomic studies, although it cannot be discarded that a virus may be involved in the onset of the disease.
- More specific analyses are required for pinpointing the aetiological agent at a finer level.

RESUMEN EN CASTELLANO

Estudio del viroma y microbioma oral asociados con la leucoplasia verrugosa proliferativa

La leucoplasia verrugosa proliferativa (LVP) es una forma maligna de leucoplasia oral (LO) que se manifiesta como parches blanquecinos hiperqueratóticos en la cavidad oral humana. Éstas son detectadas prevalentemente en mujeres de la tercera edad. La mayor parte de las LVP con el tiempo derivan en un tipo agresivo de cánceres orales, principalmente el carcinoma oral de célula escamosa (COCE).

Las asociaciones con tabaco y alcohol han sido descartadas previamente y hasta la fecha, la lesión de LVP continúa sin un agente etiológico identificado, el cual se cree pudiera ser de origen vírico debido a la alta tasa de recurrencia de la lesión y su multifocalidad (ésta reaparece en áreas distintas de la boca tras el tratamiento, el cual consiste en la remoción total del área afectada).

Con la finalidad de detectar un potencial agente causal de la lesión, se han desarrollado dos estudios independientes con biopsias de pacientes diagnosticados con LVP, LO, COCE y controles.

El primero, consistió en la búsqueda de virus oncogénicos en biopsias de pacientes usando la amplificación de reacción en cadena de la polimerasa para llevar a cabo un barrido de virus usando una batería de cebadores dirigidos a la amplificación de grupos amplios de *Herpesvirus (incluyendo los oncovirus Epstein-Bar y el virus de carcinoma de Kaposi)*, *Polyomavirus* (incluyendo el oncovirus de carcinoma de célula de Merkel) y *Papillomavirus* (incluyendo los papillomavirus de humano tipo 16 y 17 y otros identificados de alto riesgo), que incluyen algunos de los oncovirus mejor conocidos. El resultado fue negativo para todos los tipos de virus debido a que no se amplificaron fragmentos que pudieran ser identificados como pertenecientes a tales virus, con lo cual se justificó la necesidad de un segundo estudio de más amplio espectro.

Este se llevó a cabo mediante la secuenciación de la fracción bacteriana y vírica de una nueva cohorte de pacientes empleando técnicas de perfilado de la subunidad 16S del ARN ribosómico para estudiar la fracción bacteriana pertinente, así como un análisis meta genómico completo para estudiar los virus tanto de la fracción de ARN como de ADN.

El 16S del ARN ribosómico fue amplificado mediante la reacción en cadena de la polimerasa dirigido a las regiones hipervariables V1-V3 del extraído bacteriano, tras lo cual fue *pirosecuenciado* en una plataforma 454 de Roche.

Los virus de ADN y ARN fueron procesados mediante la utilización de kits comerciales para la extracción vírica, tras lo cual la fracción de virus de ADN fue enriquecida mediante un protocolo de amplificación de múltiples desplazamientos con GenomiPhi. La fracción de ARN viral fue sujeta a retrotranscripción mediante una modificación del protocolo de amplificación de cebador único independiente de secuencia. Ambas fracciones fueron secuenciadas mediante librerías de extremos pareados en plataformas MiSeq de Illumina (2x 200 en el caso del conjunto de ARN y 2x300 para los de ADN.

Con la meta de mejorar la asignación taxonómica, se desarrollaron múltiples herramientas bioinformáticas y se compilaron bases de datos de distintos dominios del árbol de la vida, específicamente creadas para la exploración del material genético secuenciado. Se decidió ampliar el estudio para también incluir la fracción de hongos, bacterias y arqueas que pudieran existir entre las secuencias de ADN y ARN secuenciada como metagenomas completos.

De igual manera, la taxonomía de la base de datos de virus fue personalizada para lidiar con problemas de asignación de rangos taxonómicos vacíos, agregando además una categoría adicional consistente en el tipo y método de replicación de los virus, la clasificación de Baltimore.

Tras un muy específico procesamiento de los datos, debido a que la amplificación de las secuencias introdujo artefactos metodológicos (particularmente en el conjunto de ARN), se lograron limpiar las secuencias y se llevó a cabo la asignación taxonómica mediante la utilización de programas y métodos de búsqueda creados para tal fin que compensan la escasa longitud de algunas secuencias utilizando un algoritmo de último ancestro común para así sólo asignar el nivel taxonómico basal cuando existe un conflicto de asignación. De tal manera, existe una mayor confianza en la asignación taxonómica ya que no se corre el riesgo de seleccionar el "mejor" resultado puesto que todos ellos son empleados y una taxonomía consenso se alcanza para cada grupo. Únicamente las secuencias que no tenían resultados en distintas tablas fueron seleccionadas.

Como resultado, se logró así identificar secuencias de cada tipo de dominio para todas las muestras. Tras el filtrado de secuencias humanas, que en este caso ocupan una gran mayoría de los datos. Una vez obtenidas tablas de contingencia, se procedió al estudio de la composición de las fracciones viral, bacteriana y fúngica, además de desarrollar la evaluación de la diversidad a nivel de muestra (diversidad alfa), así como entre ellas (diversidad beta), buscando la relación existente entre la composición y la separación por grupos, misma que no ha sido significativa para los grupos patológicos propuestos.

El análisis de biomarcadores diferencialmente abundantes no ha arrojado información soportada estadísticamente que permita definir sin lugar a dudas algún virus de la fracción de ARN o ADN como agente etiológico de las lesiones o el carcinoma. Por lo tanto, el resultado global del estudio es que no se ha podido encontrar un agente causal para la LVP.

Sin embargo, este estudio es el primero en abordar la taxonomía de hongos para estas patologías. Además, distintos marcadores no virales han sido identificados como diferencialmente distribuidos en los distintos grupos, la mayor parte de ella en la fracción bacterias. Los datos son congruentes con análisis similares de las distintas fracciones.

REFERENCES

1. J. Kallmeyer, R. Pockalny, R. R. Adhikari, D. C. Smith, S. D'Hondt, From the Cover: Global distribution of microbial abundance and biomass in subseafloor sediment. *Proc. Natl. Acad. Sci.* **109**, 16213–16216 (2012).

2. Smithsonian, Numbers of Insects (Species and Individuals). *Smithson. Inst.*, (available at http://www.si.edu/Encyclopedia_SI/nmnh/buginfo/bugnos.htm).

3. W. B. Whitman, D. C. Coleman, W. J. Wiebe, Prokaryotes: the unseen majority. *Proc. Natl. Acad. Sci.* **95**, 6578–6583 (1998).

4. K. E. Wommack, R. R. Colwell, Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* **64**, 69–114 (2000).

5. T. Fenchel, G. M. King, T. H. Blackburn, *Biogeochemistry : the Ecophysiology of Mineral Cycling* (Elsevier, London, UK, ed. 3rd, 2012).

6. K. J. Locey, J. T. Lennon, Scaling laws predict global microbial diversity. *Proc. Natl. Acad. Sci.* **Early Edit**, 1–6 (2016).

7. C. Mora, D. P. Tittensor, S. Adl, A. G. B. Simpson, B. Worm, How many species are there on earth and in the ocean? *PLoS Biol.* **9**, 1–8 (2011).

8. J. C. Lagier *et al.*, Microbial culturomics: Paradigm shift in the human gut microbiome study. *Clin. Microbiol. Infect.* **18**, 1185–1193 (2012).

9. E. J. Stewart, Growing unculturable bacteria. *J. Bacteriol.* **194**, 4151–4160 (2012).

10. N. Segata *et al.*, Computational meta'omics for microbial community studies. *Mol. Syst. Biol.* **9**, 666 (2013).

11. C. S. Pareek, R. Smoczynski, A. Tretyn, Sequencing technologies and genome sequencing. *J. Appl. Genet.* **52**, 413–435 (2011).

12. Roche, 454 Genome Sequencers (2012), (available at http://www.454.com/applications/whole-genome-sequencing/).

13. L. Wegley, R. Edwards, B. Rodriguez-Brito, H. Liu, F. Rohwer, Metagenomic analysis of the microbial community associated with the coral Porites astreoides. *Environ. Microbiol.* **9**, 2707–2719 (2007).

14. K. E. Wommack, J. Bhavsar, J. Ravel, Metagenomics: Read length matters. *Appl. Environ. Microbiol.* **74**, 1453–1463 (2008).

15. H. W. Jannasch, G. E. Jones, Bacterial populations in sea water as determined by different

methods of enumeration. *Limnol. Oceanogr.* **4**, 128–139 (1959).

16. K. H. Wilson, R. B. Blitchington, Human colonic biota studied by ribosomal DNA sequence analysis. *Appl. Environ. Microbiol.* **62**, 2273–2278 (1996).

17. F. Sanger, S. Nicklen, a R. Coulson, DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463–7 (1977).

18. C. Woese, G. Fox, Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5088–5090 (1977).

19. L. G. Wayne *et al.*, Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *Int. J. Syst. Bacteriol.* **37**, 463–464 (1987).

20. R. K. Saiki *et al.*, Primer directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*. **239**, 487–491 (1988).

21. S. J. Giovannoni, Genetic diversity in Sargasso sea bacterioplankton. *Nature*. **345**, 183–187 (1990).

22. S. J. Giovannoni, SAR11 Bacteria: The Most Abundant Plankton in the Oceans. *Ann. Rev. Mar. Sci.* **9**, 231–255 (2017).

23. R. Weller, J. W. Weller, D. M. Ward, 16S rRNA sequences of uncultivated hot spring cyanobacterial mat inhabitants retrieved as randomly primed cDNA. *Appl. Environ. Microbiol.* **57**, 1146–1151 (1991).

24. J. P. Gray, R. P. Herwig, Phylogenetic analysis of the bacterial communities in marine sediments. *Appl Env. Microbiol*. **62**, 4049–4059 (1996).

25. B. L. Maidak *et al.*, The Ribosomal Database Project (RDP-II). *Nucleic Acids Res.* **22**, 3485–3487 (1994).

26. J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy, R. M. Goodman, Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* **5**, R245–R249 (1998).

27. M. R. Rondon *et al.*, Cloning the metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.* **66**, 2541–2547 (2000).

28. O. Beja *et al.*, Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ. Microbiol.* **2**, 516–529 (2000).

29. M. Breitbart *et al.*, Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 14250–5 (2002).

30. G. W. Tyson *et al.*, Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*. **428**, 37–43 (2004).

31. J. C. Venter *et al.*, Environmental Genome Shotgun Sequencing of the. *Science.* **304**, 66–74 (2004).

32. E. F. DeLong, Microbial community genomics in the ocean. *Nat. Rev. Microbiol.* **3**, 459–69 (2005).

33. Applied Biosystems, Hitachi, ABI P RISM ® 3100 Genetic Analyzer User 's Manual (2001).

34. M. Margulies *et al.*, Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. **437**, 376–80 (2005).

35. K. Robison, Ripples from 454s Shutdown Announcement (2013), (available at http://omicsomics.blogspot.com.es/2013/10/ripples-from-454s-shutdown-announcment.html).

36. R. a Edwards *et al.*, Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics*. **7**, 57 (2006).

37. P. Yarza *et al.*, Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.* **12**, 635–645 (2014).

38. R. A. Luna *et al.*, DNA pyrosequencing-based bacterial pathogen identification in a pediatric hospital setting. *J. Clin. Microbiol.* **45**, 2985–2992 (2007).

39. S. E. Dowd *et al.*, Survey of bacterial diversity in chronic wounds using pyrosequencing, DGGE, and full ribosome shotgun sequencing. *BMC Microbiol.* **8**, 43 (2008).

40. F. E. Angly *et al.*, The marine viromes of four oceanic regions. *PLoS Biol.* **4**, 2121–2131 (2006).

41. S. J. Williamson *et al.*, The Sorcerer II Global Ocean Sampling Expedition: Metagenomic Characterization of Viruses within Aquatic Microbial Samples. *PLoS One*. **3**, e1456 (2008).

42. P. Wilmes, P. L. Bond, The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environ. Microbiol.* **6**, 911–920 (2004).

43. R. S. Poretsky *et al.*, Analysis of microbial gene transcripts in environmental samples. *Appl. Environ. Microbiol.* **71**, 4121–4126 (2005).

44. H. N. Poinar *et al.*, Metagenomics to Paleogenomics. *Science.* **311**, 392–394 (2006).

45. Solexa, Solexa 1G Genome Analysis System Brochure (2006), (available at https://www.fasteris.com/pdf/System_Profile_Brochure_10_05_06.pdf).

46. R. Cronn *et al.*, Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Res.* **36**, e122–e122 (2008).

47. Illumina, Illumina Sequencing Porfolio Brochure (2015), (available at https://www.illumina.com/content/dam/illumina-marketing/documents/products/brochures/).

48. J. Shin, G. Ming, H. Song, Decoding neural transcriptomes and epigenomes via high-throughput sequencing. *Nat. Neurosci.* **17**, 1463–1475 (2014).

49. Illumina, SBS DNA Sequencing (2014), (available at https://www.illumina.com/techniques/sequencing/dna-sequencing.html).

50. AllSeq, Sequencing Platforms.pdf (2015), (available at http://allseq.com/knowledge-bank/sequencing-platforms).

51. P. J. Turnbaugh *et al.*, A core gut microbiome in obese and lean twins. *Nature*. **457**, 480–484 (2009).

52. A. K. Bartram, M. D. J. Lynch, J. C. Stearns, G. Moreno-Hagelsieb, J. D. Neufeld, Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. *Appl. Environ. Microbiol.* **77**, 3846–3852 (2011).

53. M. Albertsen *et al.*, Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31**, 533–8 (2013).

54. T. Laver *et al.*, Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol. Detect. Quantif.* **3**, 1–8 (2015).

55. PacBio, PacBio Sequel Brochure (2017), (available at http://www.pacb.com/products-and-services/pacbio-systems/sequel/).

56. J. J. Mosher *et al.*, Improved performance of the PacBio SMRT technology for 16S rDNA sequencing. *J. Microbiol. Methods*. **104**, 59–60 (2014).

57. M. J. Sadowsky *et al.*, Analysis of Gut Microbiota – an ever changing landscape. *Gut Microbes*. **976**, 00–00 (2017).

58. S. S. Johnson, E. Zaikova, D. S. Goerlitz, Y. Bai, S. W. Tighe, Real-Time DNA Sequencing in the Antarctic Dry Valleys Using the Oxford Nanopore Sequencer. *J Biomol Tech*. **1**, 2–7 (2017).

59. S. Lax *et al.*, Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science.* **345**, 1048–1052 (2014).

60. N. A. Abreu, M. E. Taga, Decoding molecular interactions in microbial communities. *FEMS Microbiol. Rev.* **40**, 648–663 (2016).

61. S. Tighe *et al.*, Genomic Methods and Microbiological Technologies for Profiling Novel and Extreme Environments for the Extreme Microbiome Project (XMP). *J. Biomol. Tech.* **28**, 31–39 (2017).

62. R. M. Braga, M. N. Dourado, W. L. Araújo, Microbial interactions: ecology in a molecular perspective. *Brazilian J. Microbiol.* **47**, 1–13 (2016).

63. A. Reyes *et al.*, Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature*.

466, 334–338 (2010).

64. D. Rojo *et al.*, Exploring the human microbiome from multiple perspectives: factors altering its composition and function. *FEMS Microbiol. Rev.*, 1–26 (2017).

65. A. Moya, M. Ferrer, Functional Redundancy-Induced Stability of Gut Microbiota Subjected to Disturbance. *Trends Microbiol.* **24**, 402–413 (2016).

66. S. Gill, M. Pop, R. DeBoy, P. Eckburg, Metagenomic analysis of the human distal gut microbiome. *Science.* **312**, 1355–1359 (2006).

67. F. Baquero, C. Nombela, The microbiome as a human organ. *Clin. Microbiol. Infect.* **18**, 2–4 (2012).

68. A. M. O'Hara, F. Shanahan, The gut flora as a forgotten organ. *EMBO Rep*. **7**, 688–693 (2006).

69. J. Qin *et al.*, A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*. **464**, 59–65 (2010).

70. F. Backhed, Host-Bacterial Mutualism in the Human Intestine. *Science.* **307**, 1915–1920 (2005).

71. I. Ezkurdia *et al.*, Multiple evidence strands suggest that theremay be as few as 19 000 human protein-coding genes. *Hum. Mol. Genet.* **23**, 5866–5878 (2014).

72. S. A. Whiteside, H. Razvi, S. Dave, G. Reid, J. P. Burton, The microbiome of the urinary tract--a role beyond infection. *Nat. Rev. Urol.* **12**, 81–90 (2015).

73. J. G. Caporaso *et al.*, correspondence QIIME allows analysis of high- throughput community sequencing data Intensity normalization improves color calling in SOLiD sequencing. *Nat. Publ. Gr.* **7**, 335–336 (2010).

74. J. M. Marti *et al.*, Health and disease imprinted in the time variability of the human microbiome. *bioRxiv*. **2**, 10–13 (2015).

75. C. P. Tamboli, C. Neut, P. Desreumaux, J. F. Colombel, Dysbiosis in inflammatory bowel disease. *Gut*. **53**, 1–4 (2004).

76. L. V Hooper, J. I. Gordon, Commensal Host-Bacterial Relationships in the Gut .(Statistical Data Included) Commensal Host-Bacterial Relationships in the Gut .(Statistical Data Included). *Science.* **1115**, 1–7 (2001).

77. D. D. C. Savage, Microbial ecology of the gastrointestinal tract. *Annu. Rev. Microbiol.* **31**, 107–133 (1977).

78. R. Sender, S. Fuchs, R. Milo, Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol.* **14**, 1–14 (2016).

79. J. Peterson *et al.*, The NIH Human Microbiome Project. *Genome Res.* **19**, 2317–2323 (2009).

80. M. Arumugam *et al.*, Enterotypes of the human gut microbiome. *Nature*. **473**, 174–180 (2011).

81. P. J. Turnbaugh *et al.*, Feature The Human Microbiome Project. *Nature*. **449**, 804–810 (2007).

82. B. A. Methé *et al.*, A framework for human microbiome research. *Nature*. **486**, 215–221 (2012).

83. IHMP, The Integrative Human Microbiome Project: Dynamic Analysis of Microbiome-Host Omics Profiles during Periods of Human Health and Disease. *Cell Host Microbe*. **16**, 276–289 (2014).

84. D. Gevers *et al.*, The Human Microbiome Project: A Community Resource for the Healthy Human Microbiome. *PLoS Biol.* **10**, e1001377 (2012).

85. J. A. Gilbert *et al.*, Meeting report: the terabase metagenomics workshop and the vision of an Earth microbiome project. *Stand. Genomic Sci.* **3**, 243–248 (2010).

86. J. A. Gilbert, J. K. Jansson, R. Knight, The Earth Microbiome project: successes and aspirations. *BMC Biol.* **12**, 69 (2014).

87. QIITA-EMP, QIITA database statistics (2017), (available at https://qiita.ucsd.edu/stats/).

88. D. McDonald, A. Birmingham, R. Knight, Context and the human microbiome. *Microbiome*. **3**, 52 (2015).

89. The MetaSUB International Consortium, The Metagenomics and Metadesign of the Subways and Urban Biomes. *Microbiome*. **24**, 1–14 (2016).

90. E. Afshinnekoo *et al.*, Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics. *Cell Syst.* **1**, 72–87 (2015).

91. G. den Besten *et al.*, The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. *J Lipid Res*. **54**, 2325–2340 (2013).

92. G. D. Wu *et al.*, Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. *Science.* **334**, 105–108 (2011).

93. C. De Filippo *et al.*, Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 14691–6 (2010).

94. D. Knights *et al.*, Rethinking enterotypes. *Cell Host Microbe*. **16**, 433–437 (2014).

95. C. Liang *et al.*, Diversity and enterotype in gut bacterial community of adults in Taiwan. *BMC Genomics*. **18**, 932 (2017).

96. A. C. F. de Moraes *et al.*, Enterotype May Drive the Dietary-Associated Cardiometabolic Risk Factors. *Front. Cell. Infect. Microbiol.* **7**, 47 (2017).

97. C. Huttenhower *et al.*, Structure, function and diversity of the healthy human microbiome. *Nature*. **486**, 207–214 (2012).

98. National Institute of Health, NIAID Emerging Infectious Diseases / Pathogens. *Natl. Inst. Allergy Infect. Dis.* (2016), (available at https://www.niaid.nih.gov/research/emerging-infectious-diseases-pathogens).

99. J. J. Gillespie *et al.*, Patric: The comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect. Immun.* **79**, 4286–4298 (2011).

100. X. C. Morgan, N. Segata, C. Huttenhower, Biodiversity and functional genomics in the human microbiome. *Trends Genet.* **29**, 51–58 (2013).

101. K. Z. Coyte, J. Schluter, K. R. Foster, The ecology of the microbiome: Networks, competition, and stability. *Science.* **350**, 663–666 (2015).

102. A. Karkman, J. Lehtimäki, L. Ruokolainen, The ecology of human microbiota: dynamics and diversity in health and disease. *Ann. N. Y. Acad. Sci.*, 1–15 (2017).

103. K. R. Foster, T. Bell, Competition, not cooperation, dominates interactions among culturable microbial species. *Curr. Biol.* **22**, 1845–1850 (2012).

104. A. G. Loss, The Black Queen Hypothesis : Evolution of Dependencies through. **3**, 1–7 (2012).

105. N. M. Oliveira, R. Niehus, K. R. Foster, Evolutionary limits to cooperation in microbial communities. *Proc. Natl. Acad. Sci.* **111**, 201412673 (2014).

106. R. E. Ley, D. A. Peterson, J. I. Gordon, Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell.* **124**, 837–848 (2006).

107. M. Carabotti, A. Scirocco, M. A. Maselli, C. Severi, The gut-brain axis: Interactions between enteric microbiota, central and enteric nervous systems. *Ann. Gastroenterol.* **28**, 203–209 (2015).

108. M. E. Perez-Muñoz *et al.*, A critical assessment of the "sterile womb" and "in utero colonization" hypotheses: implications for research on the pioneer infant microbiome. *Microbiome*. **5**, 48 (2017).

109. E. Jiménez *et al.*, Is meconium from healthy newborns actually sterile? *Res. Microbiol.* **159**, 187–193 (2008).

110. K. Aagaard *et al.*, The Placenta Harbors a Unique Microbiome. *Sci. Transl. Med.* **6** (2014).

111. Y. Fardini, P. Chung, R. Dumm, N. Joshi, Y. W. Han, Transmission of diverse oral bacteria to murine placenta: Evidence for the oral microbiome as a potential source of intrauterine infection. *Infect. Immun.* **78**, 1789–1796 (2010).

112. H. Makino *et al.*, Mother-to-Infant Transmission of Intestinal Bifidobacterial Strains Has an Impact on the Early Development of Vaginally Delivered Infant's Microbiota. *PLoS One*. **8**, e78331 (2013).

113. M. G. Dominguez-Bello *et al.*, Delivery mode shapes the acquisition and structure of the initial

microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 11971–11975 (2010).

114. H. E. Jakobsson *et al.*, Decreased gut microbiota diversity, delayed Bacteroidetes colonisation and reduced Th1 responses in infants delivered by caesarean section. *Gut*. **63**, 559–66 (2014).

115. C. Urbaniak, M. Angelini, G. B. Gloor, G. Reid, Human milk microbiota profiles in relation to birthing method, gestation and infant gender. *Microbiome*. **4**, 1 (2016).

116. D. M. Chu *et al.*, Maturation of the infant microbiome community structure and function across multiple body sites and in relation to mode of delivery. *Nat. Med.* (2017).

117. H. Chu, S. K. Mazmanian, Innate immune recognition of the microbiota promotes host-microbial symbiosis. *Nat. Immunol.* **14**, 668–75 (2013).

118. C. Palmer, E. M. Bik, D. B. DiGiulio, D. A. Relman, P. O. Brown, Development of the human infant intestinal microbiota. *PLoS Biol.* **5**, 1556–1573 (2007).

119. M. J. Claesson *et al.*, Gut microbiota composition correlates with diet and health in the elderly. *Nature*. **488**, 178–184 (2012).

120. N. Ottman, H. Smidt, W. M. de Vos, C. Belzer, The function of our microbiota: who is out there and what do they do? *Front Cell Infect Microbiol*. **2**, 104 (2012).

121. E. A. Grice, J. A. Segre, The skin microbiome. *Nat. Rev. Microbiol.* **9**, 244–53 (2011).

122. L. M. Collins, C. Dawes, The surface area of the adult human mouth and thickness of the salivary film covering the teeth and oral mucosa. *J. Dent. Res.* **66**, 1300–2 (1987).

123. C. Iida-Kondo *et al.*, Comparison of tongue volume/oral cavity volume ratio between obstructive sleep apnea syndrome patients and normal adults using magnetic resonance imaging. *J. Med. Dent. Sci.* **53**, 119–126 (2006).

124. F. E. Dewhirst *et al.*, The human oral microbiome. *J. Bacteriol.* **192**, 5002–5017 (2010).

125. S. Standring, *Gray's Anatomy - The Anatomical Basis of Clinical Practice* (Elsevier, London, UK, ed. 41, 2016).

126. A. Nanci, *Ten Cate's Oral Histology Development, Structure and Function* (Elsevier Mosby, St. Louis, MO, ed. 8, 2013).

127. P. Williams, M. Marsh, D. Lewis, *Oral Microbiology* (Churchill Livingstone, Elsevier, Lomdon, UK, ed. 5, 2009).

128. S. S. Socransky, A. D. Haffajee, Dental biofilms: difficult therapeutic targets. *Periodontol. 2000*. **28**, 12–55 (2002).

129. R. M. Donlan, J. W. Costerton, Biofilms: survivalmechanisms of clinically relevant microorganisms. *Clin.Microbiol. Rev.* **15**, 167–19 (2002).

130. T. Rosebury, The Challenge to Dentistry. **126** (1957).

131. T. Rosebury, L. Waugh, Dental caries among Eskimos of the Kuskokwim area of Alaska. *Am. J. Dis. Child.* **57**, 871 (1939).

132. J. K. Clarke, On the bacterial factor in the aetiology of dental caries. *Br. J. Exp. Pathol.* **5**, 141–147 (1924).

133. T. Rosebury, G. Foley, S. Greenberg, Studies of Lactobacilli in relation to caries in rats. II. Attempt to immunize rats, on cariesproducing diets, against Lactobacilli. *Jorunal Dent. Res.*, 231–232 (1934).

134. T. Rosebury, G. Foley, Experimental Vincent's Infection. *J. Am. Dent. Assoc.* **26**, 1798–1811 (1939).

135. T. Rosebury, Recent developments in fuso-spirochetal and related infections: the role of infection in Vincent's and related diseases. *J Periodontol*. **17**, 121–125 (1946).

136. G. Foley, T. Rosebury, Comparative Infectivity for Guinea Pigs of Fuso-Spirochetal Exudates from Different Diseases. *J. Dent. Res.* **21**, 375–378 (1942).

137. T. Rosebury, A. R. Clark, J. B. MacDonald, D. C. O'Connell, Studies of Fusospirochetal Infection: III. Further Studies of a Guinea Pig Passage Strain of Fusospirochetal Infection, Including the Infectivity of Sterile Exudate Filtrates, of Mixed Cultures Through Ten Transfers, and of Recombined Pure Cultures. *J. Infect. Dis.* **87**, 234–248 (1950).

138. T. Rosebury, J. B. Reynolds, Continuous anaerobiosis for cultivation of spirochetes. *Proc. Soc. Exp. Biol. Med.* **117**, 813–5 (1964).

139. D. H. Fine, Dr. Theodor Rosebury: Grandfather of Modern Oral Microbiology. *J. Dent. Res.* **85**, 990–995 (2006).

140. R. J. Fitzgerald, P. H. Keyes, Demonstration of the etiologic role of streptococci in experimental caries in the hamster. *J. Am. Dent. Assoc.* **61**, 9–19 (1960).

141. A. Jepsen, J. E. Winther, Mycotic Infection in Oral Leukoplakia. *Acta Odontol. Scand.* **23**, 239–256 (1965).

142. H. I. Winner, The Transition From Commensalism To Parasitism. *Br. J. Dermatol.* **81**, 62–68 (1969).

143. C. Russell, J. H. Jones, Effects of oral inoculation of Candida albicans in tetracycline-treated rats. *J. Med. Microbiol.* **6** (1973).

144. R. P. Teles *et al.*, Rediscovering Sig Socransky, the genius and his legacy. *J. Dent. Res.* **91**, 433–9 (2012).

145. S. S. Socransky, W. J. Loesche, C. Hubersak, J. B. MacDonald, Dependency of treponema

microdentium on other oral organisms for isobutyrate, polyamines, and a controlled oxidation-reduction potential. *J. Bacteriol.* **88**, 200–9 (1964).

146. M. A. Listgarten, S. S. Socransky, Electron microscopy of axial fibrils, outer envelope, and cell division of certain oral spirochetes. *J. Bacteriol.* **88**, 1087–103 (1964).

147. W. J. Loesche, R. J. Gibbons, S. S. Socransky, Biochemical characteristics of Vibrio sputorum and relationship to Vibrio bubulus and Vibrio fetus. *J. Bacteriol.* **89**, 1109–16 (1965).

148. W. J. Loesche, S. S. Socransky, R. J. Gibbons, Bacteroides oralis, proposed new species isolated from the oral cavity of man. *J. Bacteriol.* **88**, 1329–37 (1964).

149. S. S. Socransky, M. Listgarten, C. Hubersak, J. Cotmore, A. Clark, Morphological and biochemical differentiation of three types of small oral spirochetes. *J. Bacteriol.* **98**, 878–882 (1969).

150. E. R. Leadbetter, S. C. Holt, S. S. Socransky, Capnocytophaga: New genus of gram-negative gliding bacteria I. General characteristics, taxonomic considerations and significance. *Arch. Microbiol.* **122**, 9–16 (1979).

151. J. De Ley, H. Cattoir, a Reynaerts, The quantitative measurement of DNA hybridization from renaturation rates. *Eur. J. Biochem.* **12**, 133–142 (1970).

152. M. Kilian, A Taxonomic Study of the Genus Haemophilus, with the Proposal of a New Species. *J. Gen. Microbiol.* **93**, 9–62 (1976).

153. S. K. P. Lau *et al.*, Characterization of Haemophilus segnis , an important cause of bacteremia, by 16S rRNA gene sequencing. *J. Clin. Microbiol.* **42**, 877–80 (2004).

154. A. C. R. Tanner *et al.*, Wolinella gen. nov., Wolinella succinogenes (Vibrio succinogenes Wolin et al.) comb. nov., and Description of Bacteroides gracilis sp. nov., Wolinella recta sp. nov., Campylobacter concisus sp. nov., and Eikenella corrodens from Humans with Periodontal Di. *Int. J. Syst. Bacteriol.* **31**, 432–445 (1981).

155. A. C. R. Tanner, R. A. Visconti, S. S. Socransky, S. C. Holt, Classification and identification of Actinobacillus actinomycetemcomitans and Haemophilus aphrophilus by cluster analysis and deoxyribonucleic acid hybridizations. *J. Periodontal Res.* **17**, 585–596 (1982).

156. M. A. Listgarten, Structure of the Microbial Flora Associated with Periodontal Health and Disease in Man: A Light and Electron Microscopic Study. *J. Periodontol.* **47**, 1–18 (1976).

157. J. W. Costerton, G. G. Geesey, K. J. Cheng, How bacteria stick. *Sci. Am.* **238**, 86–95 (1978).

158. J. W. Costerton *et al.*, Bacterial Biofilms in Nature and Disease. *Annu. Rev. Microbiol.* **41**, 435–464 (1987).

159. S. S. Socransky *et al.*, “Checkerboard” DNA-DNA hybridization. *Biotechniques*. **17**, 788–792 (1994).

160. S. S. Socransky, a D. Haffajee, M. a Cugini, C. Smith, R. L. Kent, Microbial complexes in subgingival plaque. *J. Clin. Periodontol.* **25**, 134–144 (1998).

161. A. Tanner, E. Dewhirst, The impact of 16S ribosomal RNA-based phylogeny on the taxonomy of oral bacteria. *Periodontol. 2000*. **5**, 26–51 (2000).

162. H. Thompson, A. Rybalka, R. Moazzez, F. E. Dewhirst, W. G. Wade, In vitro culture of previously uncultured oral bacterial phylotypes. *Appl. Environ. Microbiol.* **81**, 8307–8314 (2015).

163. Forsyth Institute, HOMD - Human Oral Microbiome Database (2017), (available at http://www.homd.org/index.php?name=HOMD&taxonomy_level=1).

164. B. J. Paster *et al.*, Bacterial Diversity in Human Subgingival Plaque. *J. Bacteriol.* **183**, 3770–3783 (2001).

165. M. Faveri *et al.*, Microbiological diversity of generalized aggressive periodontitis by 16S rRNA clonal analysis. *Oral Microbiol. Immunol.* **23**, 112–118 (2008).

166. B. J. Paster *et al.*, Prevalent bacterial species and novel phylotypes in advanced noma lesions. *J. Clin. Microbiol.* **40**, 2187–2191 (2002).

167. M. a Munson, T. Pitt-Ford, B. Chong, a Weightman, W. G. Wade, Molecular and cultural analysis of the microflora associated with endodontic infections. *J. Dent. Res.* **81**, 761–766 (2002).

168. B. J. Paster *et al.*, Bacterial diversity in necrotizing ulcerative periodontitis in HIV-positive subjects. *Ann. Periodontol.* **7**, 8–16 (2002).

169. M. R. Becker *et al.*, Molecular analysis of bacterial species associated with childhood caries. *J Clin Microbiol*. **40**, 1001–1009 (2002).

170. C. E. Kazor *et al.*, Diversity of bacterial populations on the tongue dorsa of patients with halitosis and healthy patients. *J. Clin. Microbiol.* **41**, 558–563 (2003).

171. M. A. M. Munson, A. Banerjee, T. F. Watson, W. G. Wade, Molecular analysis of the microflora associated with dental caries. *J. Clin. …*. **42**, 3023–9 (2004).

172. A. Ferrari, T. Brusa, A. Rutili, E. Canzi, B. Biavati, Isolation and characterization of Methanobrevibacter oralis sp. nov. *Curr. Microbiol.* **29**, 7–12 (1994).

173. P. W. Lepp *et al.*, Methanogenic Archaea and human periodontal disease. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 6176–81 (2004).

174. F. Matarazzo, A. C. Ribeiro, M. Feres, M. Faveri, M. P. A. Mayer, Diversity and quantitative analysis of Archaea in aggressive periodontitis and periodontally healthy subjects. *J. Clin. Periodontol.* **38**, 621–627 (2011).

175. J. a Aas, B. J. Paster, L. N. Stokes, I. Olsen, F. E. Dewhirst, Defining the Normal Bacterial Flora of the Oral Cavity. *J. Clin. Microbiol.* **43**, 5721–5732 (2005).

176. D. L. Mager, L. A. Ximenez-Fyvie, A. D. Haffajee, S. S. Socransky, Distribution of selected bacterial species on intraoral surfaces. *J. Clin. Periodontol.* **30**, 644–54 (2003).

177. P. M. Corby *et al.*, Heritability of Oral Microbial Species in Caries-Active and Caries-Free Twins. *Twin Res. Hum. Genet.* **10**, 821–828 (2007).

178. T. Chen *et al.*, The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database*. **2010** (2010).

179. S. R. Vartoukian, R. M. Palmer, W. G. Wade, Diversity and morphology of members of the phylum "Synergistetes" in periodontal health and disease. *Appl. Environ. Microbiol.* **75**, 3777–3786 (2009).

180. A. P. Colombo *et al.*, Comparisons of subgingival microbial profiles of refractory periodontitis, severe periodontitis, and periodontal health using the human oral microbe identification microarray. *J. Periodontol.* **80**, 1132–1421 (2009).

181. A. P. V Colombo *et al.*, Impact of periodontal therapy on the subgingival microbiota of severe periodontitis: comparison between good responders and individuals with refractory periodontitis using the human oral microbe identification microarray. *J. Periodontol.* **83**, 1279–87 (2012).

182. Forsyth Institute, HOMINGS: Human Oral Microbe Identification using Next G eneration Sequencing - Species level identification of nearly 600 oral bacterial taxa (2017), (available at http://homings.forsyth.org/index2.html).

183. B. J. F. Keijser *et al.*, Pyrosequencing analysis of the oral microflora of healthy adults. *J. Dent. Res.* **87**, 1016–1020 (2008).

184. E. Zaura, B. J. F. Keijser, S. M. Huse, W. Crielaard, Defining the healthy core microbiome of oral microbial communities. *BMC Microbiol.* **9**, 259 (2009).

185. V. Lazarevic *et al.*, Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *J. Microbiol. Methods*. **79**, 266–271 (2009).

186. N. Segata *et al.*, Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biol.* **13**, R42 (2012).

187. V. Lazarevic, K. Whiteson, D. Hernandez, P. Francois, J. Schrenzel, Study of inter- and intra-individual variations in the salivary microbiota. *BMC Genomics*. **11**, 523 (2010).

188. E. M. Bik *et al.*, Bacterial diversity in the oral cavity of 10 healthy individuals. *ISME J.* **4**, 962–974 (2010).

189. L. Abusleme *et al.*, The subgingival microbiome in health and periodontitis and its relationship with community biomass and inflammation. *Isme J.* **7**, 1016–1025 (2013).

190. A. L. Griffen *et al.*, Distinct and complex bacterial profiles in human periodontitis and health revealed by 16S pyrosequencing. *ISME J.* **6**, 1176–1185 (2011).

191. S. Bizzarro, B. G. Loos, M. L. Laine, W. Crielaard, E. Zaura, Subgingival microbiome in smokers and non-smokers in periodontitis: An exploratory study using traditional targeted techniques and a next-generation sequencing. *J. Clin. Periodontol.* **40**, 483–492 (2013).

192. P. Belda-Ferre *et al.*, The oral metagenome in health and disease. *ISME J.* **6**, 46–56 (2012).

193. A. Camelo-Castillo, A. Benítez-Pérez, P. Belda-Ferre, R. Cabrera-Rubio, A. Mira, Streptococcus dentisani sp. nov., a novel member of the mitis group. *Int. J. Syst. Evol. Microbiol.* **64**, 60–65 (2014).

194. A. López-López, A. Camelo-Castillo, M. D. Ferrer, Á. Simon-Soro, A. Mira, Health-Associated Niche Inhabitants as Oral Probiotics: The Case of Streptococcus dentisani. *Front. Microbiol.* **8**, 1–12 (2017).

195. N. Takahashi, B. Nyvad, The role of bacteria in the caries process: ecological perspectives. *J. Dent. Res.* **90**, 294–303 (2011).

196. Y. H. Li, X. Tian, Quorum sensing and bacterial social interactions in biofilms. *Sensors*. **12**, 2519–2538 (2012).

197. D. A. Devine, P. D. Marsh, J. Meade, Modulation of host responses by oral commensal bacteria. *J. Oral Microbiol.* **7**, 26941 (2015).

198. P. D. Marsh, D. A. Head, D. A. Devine, Ecological approaches to oral biofilms: Control without killing. *Caries Res.* **49**, 46–54 (2015).

199. J. L. Mark Welch, B. J. Rossetti, C. W. Rieken, F. E. Dewhirst, G. G. Borisy, Biogeography of a human oral microbiome at the micron scale. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E791-800 (2016).

200. R. J. Lamont, G. N. Hajishengallis, H. F. Jenkinson, Oral Microbiology and Immunology, Second Edition, 2014 (2014).

201. I. R. Kramer, R. B. Lucas, J. J. Pindborg, L. H. Sobin, Definition of leukoplakia and related lesions: an aid to studies on oral precancer. *Oral Surg. Oral Med. Oral Pathol.* **46**, 518–539 (1978).

202. S. Petti, Pooled estimate of world leukoplakia prevalence: A systematic review. *Oral Oncol.* **39**, 770–780 (2003).

203. F. S. Mehta, J. J. Pindborg, P. C. Gupta, D. K. Daftary, Odont, Epidemiologic and histologic study of oral cancer and leukoplakia among 50,915 villagers in India. *Cancer*. **24**, 832–849 (1969).

204. S. S. Napier, P. M. Speight, Natural history of potentially malignant oral lesions and conditions:

An overview of the literature. *J. Oral Pathol. Med.* **37**, 1–10 (2008).

205. A. Villa, S. Bin Woo, Leukoplakia—A Diagnostic and Management Algorithm. *J. Oral Maxillofac. Surg.* **75**, 723–734 (2017).

206. G. Lodi, A. Sardella, C. Bez, F. Demarosi, A. Carrassi, *Interventions for treating oral leukoplakia* (John Wiley & Sons, Ltd, Chichester, UK, 2006).

207. L. S. Hansen, J. A. Olson, S. Silverman, Proliferative verrucous leukoplakia. A long-term study of thirty patients. *Oral Surg. Oral Med. Oral Pathol.* **60**, 285–98 (1985).

208. S. Silverman, M. Gorsky, F. Lozada, Oral Leukoplakia and Malignant Transformation.A follow-up study of 257 patients. *Cancer*. **53**, 563–568 (1984).

209. C. E. Vigliante, P. D. Quinn, F. Alawi, Proliferative verrucous leukoplakia: Report of a case with characteristic long-term progression. *J. Oral Maxillofac. Surg.* **61**, 626–631 (2003).

210. R. García-López, A. Moya, J. V. Bagan, V. Pérez-Brocal, Retrospective case-control study of viral pathogen screening in proliferative verrucous leukoplakia lesions. *Clin. Otolaryngol.* **39**, 272–280 (2014).

211. R. Cerero-Lapiedra, D. Balade-Martinez, L. Moreno-Lopez, G. Esparza-Gomez, J. Bagan, Proliferative verrucous leukoplakia: A proposal for diagnostic criteria. *Med. Oral Patol. Oral y Cir. Bucal*. **15**, e839–e845 (2010).

212. J. M. Zakrzewska, V. Lopes, P. Speight, C. Hopper, Proliferative verrucous leukoplakia: a report of ten cases. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.* **82**, 396–401 (1996).

213. J. V Bagan, R. Poveda, M. A. Milian, J. Murillo, Proliferative verrucous leukoplakia : high incidence of gingival squamous cell carcinoma. **1985**, 379–382 (2003).

214. J. Bagan, C. Scully, Y. Jimenez, M. Martorell, Proliferative verrucous leukoplakia: a concise update. *Oral Dis.* **16**, 328–32 (2010).

215. M. Pentenero, M. Meleti, P. Vescovi, S. Gandolfo, Oral proliferative verrucous leucoplakia: Are there particular features for such an ambiguous entity? A systematic review. *Br. J. Dermatol.* **170**, 1039–1047 (2014).

216. S. Silverman Jr, M. Gorsky, Proliferative verrucous leukoplakia: A follow-up study of 54 cases. *Oral Surgery, Oral Med. Oral Pathol. Oral Radiol. Endodontology*. **84**, 154–157 (1997).

217. A. Fettig *et al.*, Proliferative verrucous leukoplakia of the gingiva. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.* **90**, 723–730 (2000).

218. J. V Bagán *et al.*, Proliferative verrucous leukoplakia: unusual locations of oral squamous cell carcinomas, and field cancerization as shown by the appearance of multiple OSCCs. *Oral Oncol.* **40**, 440–3 (2004).

219. S. Gandolfo, R. Castellani, M. Pentenero, Proliferative verrucous leukoplakia: a potentially malignant disorder involving periodontal sites. *J. Periodontol.* **80**, 274–81 (2009).

220. U. Romeo, N. Russo, G. Palaia, G. Tenore, A. Del Vecchio, Oral proliferative verrucous leukoplakia treated with the photodynamic therapy: a case report. *Ann. Stomatol. (Roma).* **5**, 77–80 (2014).

221. A. Martinez-Lopez *et al.*, Successful treatment of proliferative verrucous leukoplakia with 5% topical imiquimod. *Dermatol. Ther.* **30**, e12413 (2017).

222. C. B. Roman, H. O. Sedano, Multifocal papilloma virus epithelial hyperplasia. *Oral Surgery, Oral Med. Oral Pathol.* **77**, 631–635 (1994).

223. J. M. Palefsky, S. J. Silverman, M. Abdel-Salaam, T. E. Daniels, J. S. Greenspan, Association between proliferative verrucous leukoplakia and infection with human papillomavirus type 16. *J. Oral Pathol. Med.* **24**, 193–197 (1995).

224. R. Gopalakrishnan *et al.*, Mutated and wild-type p53 expression and HPV integration in proliferative verrucous leukoplakia and oral squamous cell carcinoma. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.* **83**, 471–477 (1997).

225. J. V Bagan *et al.*, Lack of Association Between Proliferative Verrucous Leukoplakia and Human Papillomavirus Infection. *J. Oral Maxillofac. Surg.* **65**, 46–49 (2007).

226. J. S. Greenspan, D. Greenspan, J. Webster-Cyriaque, Hairy leukoplakia; lessons learned: 30-plus years. *Oral Dis.* **22**, 120–127 (2016).

227. J. V. Bagan *et al.*, Epstein-Barr virus in oral proliferative verrucous leukoplakia and squamous cell carcinoma: A preliminary study. *Med. Oral Patol. Oral Cir. Bucal*. **13**, 110–113 (2008).

228. M. Plummer *et al.*, Global burden of cancers attributable to infections in 2012: a synthetic analysis. *Lancet Glob. Heal.* **4**, e609–e616 (2016).

229. W. L. Dissanayaka *et al.*, Clinical and histopathologic parameters in survival of oral squamous cell carcinoma. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* **113**, 518–525 (2012).

230. D. W. Jung *et al.*, Tumor-stromal crosstalk in invasion of oral squamous cell carcinoma: A pivotal role of CCL7. *Int. J. Cancer*. **127**, 332–344 (2010).

231. R. L. Siegel, K. D. Miller, A. Jemal, Cancer statistics, 2017. *CA. Cancer J. Clin.* **67**, 7–30 (2017).

232. O. Bettendorf, J. Piffkò, A. Bànkfalvi, Prognostic and predictive factors in oral squamous cell cancer: Important tools for planning individual therapy? *Oral Oncol.* **40**, 110–119 (2004).

233. A. Celetti *et al.*, *Intraepithelial Neoplasia* (InTech, 2012).

234. D. Hanahan, R. A. Weinberg, The hallmarks of cancer. *Cell*. **100**, 57–70 (2000).

235. J. Califano, W. H. Westra, G. Meininger, R. Corio, W. M. Koch, Advances in Brief Genetic Progression and Clonal Relationship of Recurrent Premalignant Head and Neck Lesions. **6**, 347–352 (2000).

236. C. R. Leemans, B. J. M. Braakhuis, R. H. Brakenhoff, The molecular biology of head and neck cancer. *Nat. Rev. Cancer*. **11**, 9–22 (2011).

237. C. C. Bitu *et al.*, HOXB7 expression is a prognostic factor for oral squamous cell carcinoma. *Histopathology*. **60**, 662–665 (2012).

238. J. Massano, F. S. Regateiro, G. Januário, A. Ferreira, Oral squamous cell carcinoma: Review of prognostic and predictive factors. *Oral Surgery, Oral Med. Oral Pathol. Oral Radiol. Endodontology*. **102**, 67–76 (2006).

239. D. Wangsa *et al.*, Ki-67 expression predicts locoregional recurrence in stage I oral tongue carcinoma. *Br. J. Cancer*. **99**, 1121–1128 (2008).

240. H.-X. Fan, H.-X. Li, D. Chen, Z.-X. Gao, J.-H. Zheng, Changes in the expression of MMP2, MMP9, and ColIV in stromal cells in oral squamous tongue cell carcinoma: relationships and prognostic implications. *J. Exp. Clin. Cancer Res.* **31**, 90 (2012).

241. A. Zaravinos, An updated overview of HPV-associated head and neck carcinomas. *Oncotarget*. **5**, 3956–69 (2014).

242. S. Pushalkar *et al.*, Microbial diversity in saliva of oral squamous cell carcinoma. *FEMS Immunol. Med. Microbiol.* **61**, 269–277 (2011).

243. S. Pushalkar *et al.*, Comparison of oral microbiota in tumor and non-tumor tissues of patients with oral squamous cell carcinoma. *BMC Microbiol.* **12**, 144 (2012).

244. D. A. Baldwin, M. Feldman, J. C. Alwine, E. S. Robertson, Metagenomic assay for identification of microbial pathogens in tumor tissues. *MBio*. **5**, e01714-14 (2014).

245. N. N. Al-hebshi *et al.*, Inflammatory bacteriome featuring Fusobacterium nucleatum and Pseudomonas aeruginosa identified in association with oral squamous cell carcinoma. *Sci. Rep.* **7**, 1834 (2017).

246. N. Popgeorgiev *et al.*, Marseillevirus-Like Virus Recovered From Blood Donated by Asymptomatic Humans. *J. Infect. Dis.* **208**, 1042–1050 (2013).

247. S. Spandole, D. Cimponeriu, L. M. Berca, G. Mihăescu, Human anelloviruses: an update of molecular, epidemiological and clinical aspects. *Arch. Virol.* **160**, 893–908 (2015).

248. A. Moustafa *et al.*, The blood DNA virome in 8,000 humans. *PLOS Pathog.* **13**, e1006292 (2017).

249. T. J. Meyer, J. L. Rosenkrantz, L. Carbone, S. L. Chavez, Endogenous Retroviruses: With Us and against Us. *Front. Chem.* **5**, 1–8 (2017).

250. J. S. Flint, L. W. Enquist, V. R. Racaniello, G. F. Rall, A. M. Skalka, *Principles of Virology* (American Society of Microbiology, ed. 4th, 2015).

251. P. Aiewsakun, A. Katzourakis, Marine origin of retroviruses in the early Palaeozoic Era. *Nat. Commun.* **8**, 13954 (2017).

252. D. R. Lowy, W. P. Rowe, N. Teich, J. W. Hartley, Murine leukemia virus: high-frequency activation in vitro by 5-iododeoxyuridine and 5-bromodeoxyuridine. *Science*. **174**, 155–6 (1971).

253. J. Kool, A. Berns, High-throughput insertional mutagenesis screens in mice to identify oncogenic networks. *Nat. Rev. Cancer*. **9**, 389–399 (2009).

254. A. Dupressoir *et al.*, A pair of co-opted retroviral envelope syncytin genes is required for formation of the two-layered murine placental syncytiotrophoblast. *Proc. Natl. Acad. Sci.* **108**, E1164–E1173 (2011).

255. John F. Atkins, R. F. Gesteland, *Recoding: Expansion of Decoding Rules Enriches Gene Expression* (Springer-verlag, New York, USA, 2010).

256. J. H. Paul, M. B. Sullivan, A. M. Segall, F. Rohwer, Marine phage genomics. *Comp. Biochem. Physiol. - B Biochem. Mol. Biol.* **133**, 463–476 (2002).

257. L. Deng *et al.*, Viral tagging reveals discrete populations in Synechococcus viral genome sequence space. *Nature*. **513**, 242–245 (2014).

258. International Committee on Taxonomy of Viruses, ICTV Species List 2016 v 1.2, (available at https://talk.ictvonline.org/).

259. A. J. Gibbs, Viral taxonomy needs a spring clean. Its exploration era is over. *Virol. J.* **10**, 254 (2013).

260. M. H. V van Regenmortel *et al.*, Virus species polemics: 14 senior virologists oppose a proposed change to the ICTV definition of virus species. *Arch. Virol.* **158**, 1115–1119 (2013).

261. G. J. Morgan, What is a virus species? Radical pluralism in viral taxonomy. *Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.* **59**, 64–70 (2016).

262. C. H. Calisher, The taxonomy of viruses should include viruses. *Arch. Virol.* **161**, 1419–1422 (2016).

263. N. Beerenwinkel, H. F. Günthard, V. Roth, K. J. Metzner, Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Front. Microbiol.* **3**, 1–16 (2012).

264. V. Pérez-Brocal *et al.*, Study of the viral and microbial communities associated with Crohn's disease: a metagenomic approach. *Clin. Transl. Gastroenterol.* **4**, e36 (2013).

265. F. Rohwer, M. Youle, H. Maughan, N. Hisakawa, *Life in Our Phage World* (Wholon, San Diego,

USA, ed. 1, 2015).

266. M. Breitbart, F. Rohwer, Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. *Biotechniques*. **39**, 729–736 (2005).

267. R. a. Edwards, F. Rohwer, Opinion: Viral metagenomics. *Nat. Rev. Microbiol.* **3**, 504–510 (2005).

268. D. Moreira, P. López-García, Ten reasons to exclude viruses from the tree of life. *Nat. Rev. Microbiol.* **7**, 306–311 (2009).

269. F. Rohwer, R. Edwards, The phage proteomic tree: A genome-based taxonomy for phage. *J. Bacteriol.* **184**, 4529–4535 (2002).

270. C. Proux *et al.*, The dilemma of phage taxonomy illustrated by comparative genomics of Sfi21-like Siphoviridae in lactic acid bacteria. *J. Bacteriol.* **184**, 6026–36 (2002).

271. J. G. Lawrence, G. F. Hatfull, R. W. Hendrix, Imbroglios of viral taxonomy: Genetic exchange and failings of phenetic approaches. *J. Bacteriol.* **184**, 4891–4905 (2002).

272. D. H. Bamford, Do viruses form lineages across different domains of life? *Res. Microbiol.* **154**, 231–236 (2003).

273. S. Yooseph *et al.*, The Sorcerer II global ocean sampling expedition: Expanding the universe of protein families. *PLoS Biol.* **5**, 0432–0466 (2007).

274. J. Li, S. K. Halgamuge, S.-L. Tang, Genome classification by gene distribution: an overlapping subspace clustering approach. *BMC Evol. Biol.* **8**, 116 (2008).

275. A. L. Grazziotin, E. V Koonin, D. M. Kristensen, Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res.* **45**, 491–498 (2017).

276. E. V. Koonin, V. V. Dolja, M. Krupovic, Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology*. **479–480**, 2–25 (2015).

277. J. L. Goodier, H. H. Kazazian, Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell*. **135**, 23–35 (2008).

278. N. Yutin, D. Raoult, E. V Koonin, Virophages, polintons, and transpovirons: a complex evolutionary network of diverse selfish genetic elements with different reproduction strategies. *Virol. J.* **10**, 158 (2013).

279. T. Zhang *et al.*, RNA viral community in human feces: Prevalence of plant pathogenic viruses. *PLoS Biol.* **4**, 0108–0118 (2006).

280. F. Balique, P. Colson, D. Raoult, Tobacco mosaic virus in cigarettes and saliva of smokers. *J. Clin. Virol.* **55**, 374–376 (2012).

281. P. Colson *et al.*, Pepper mild mottle virus, a plant virus associated with specific immune responses, fever, abdominal pains, and pruritus in humans. *PLoS One*. **5** (2010), doi:10.1371/journal.pone.0010041.

282. D. Willner *et al.*, Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl**, 4547–53 (2011).

283. D. T. Pride *et al.*, Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. *ISME J.* **6**, 915–26 (2012).

284. R. Robles-Sikisaka *et al.*, Association between living environment and human oral viral ecology. *ISME J.* **7**, 1710–24 (2013).

285. S. R. Abeles *et al.*, Human oral viruses are personal, persistent and gender-consistent. *ISME J.* **8**, 1–15 (2014).

286. M. Ly *et al.*, Altered oral viral ecology in association with periodontal disease. *MBio*. **5** (2014), doi:10.1128/mBio.01133-14.

287. E. Delwart, L. Li, Rapidly expanding genetic diversity and host range of the Circoviridae viral family and other Rep encoding small circular ssDNA genomes. *Virus Res.* **164**, 114–121 (2012).

288. J. A. Vanchiere *et al.*, Polyomavirus shedding in the stool of healthy adults. *J. Clin. Microbiol.* **47**, 2388–2391 (2009).

289. D. M. P. G. Barreira *et al.*, Viral load and genotypes of noroviruses in symptomatic and asymptomatic children in Southeastern Brazil. *J. Clin. Virol.* **47**, 60–64 (2010).

290. K. Thom, C. Morrison, J. C. M. Lewis, P. Simmonds, Distribution of TT virus (TTV), TTV-like minivirus, and related viruses in humans and nonhuman primates. *Virology*. **306**, 324–333 (2003).

291. K. M. Wylie, K. a. Mihindukulasuriya, E. Sodergren, G. M. Weinstock, G. a. Storch, Sequence analysis of the human virome in Febrile and Afebrile children. *PLoS One*. **7** (2012), doi:10.1371/journal.pone.0027735.

292. N. Popgeorgiev, S. Temmam, D. Raoult, C. Desnues, Describing the silent human virome with an emphasis on giant viruses. *Intervirology*. **56**, 395–412 (2013).

293. D. Paez-Espino *et al.*, Uncovering Earth’s virome. *Nature*. **536**, 425–430 (2016).

294. A. Simón-Soro, M. Guillen-Navarro, A. Mira, Metatranscriptomics reveals overall active bacterial composition in caries lesions. *J. Oral Microbiol.* **6**, 1–6 (2014).

295. J. C. Wooley, A. Godzik, I. Friedberg, A Primer on Metagenomics. *PLoS Comput. Biol.* **6**, e1000667 (2010).

296. A. Reyes, M. Wu, N. P. McNulty, F. L. Rohwer, J. I. Gordon, Gnotobiotic mouse model of

phage-bacterial host dynamics in the human gut. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 20236–41 (2013).

297. P. J. Turnbaugh *et al.*, The Effect of Diet on the Human Gut Microbiome: A Metagenomic Analysis in Humanized Gnotobiotic Mice. *Sci. Transl. Med.* **1**, 1–10 (2009).

298. J. F. Vázquez-Castellanos, R. García-López, V. Pérez-Brocal, M. Pignatelli, A. Moya, Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. *BMC Genomics*. **15**, 37 (2014).

299. W. A. Walters *et al.*, PrimerProspector: De novo design and taxonomic analysis of barcoded polymerase chain reaction primers. *Bioinformatics*. **27**, 1159–1161 (2011).

300. O. Forslund, A. Antonsson, P. Nordin, B. Stenquist, B. G. Hansson, A broad range of human papillomavirus types detected with a general PCR method suitable for analysis of cutaneous tumours and normal skin. *J. Gen. Virol.* **80**, 2437–2443 (1999).

301. A. Iftner *et al.*, The prevalence of human papillomavirus genotypes in nonmelanoma skin cancers of nonimmunosuppressed individuals identifies high-risk genital types as possible risk factors. *Cancer Res.* **63**, 7515–9 (2003).

302. F. H. Leendertz *et al.*, African Great Apes Are Naturally Infected with Polyomaviruses Closely Related to Merkel Cell Polyomavirus. *J. Virol.* **85**, 916–924 (2011).

303. E. J. Duncavage, B. A. Zehnbauer, J. D. Pfeifer, Prevalence of Merkel cell polyomavirus in Merkel cell carcinoma. *Mod. Pathol.* **22**, 516–21 (2009).

304. D. R. Vandevanter *et al.*, Detection and analysis of diverse herpesviral species by consensus primer PCR. *J. Clin. Microbiol.* **34**, 1666–1671 (1996).

305. J. K. Bonfield, K. f Smith, R. Staden, A new DNA sequence assembly program. *Nucleic Acids Res.* **23**, 4992–4999 (1995).

306. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–10 (1990).

307. G. R. Reyes, J. P. Kim, Sequence-independent, single-primer amplification (SISPA) of complex DNA populations. *Mol. Cell. Probes*. **5**, 473–481 (1991).

308. A. Djikeng *et al.*, Viral genome sequencing by random priming methods. *BMC Genomics*. **9**, 5 (2008).

309. F. Dean, J. Nelson, T. Giesler, R. Lasken, Rapid amplification of plasmid and phage DNA using Phi29 polymerase and a multiply-pimed rolling circle amplification. *Genome Res.* **11**, 1095–1099 ST–Rapid amplification of plasmid a (2001).

310. S. Balzer, K. Malde, I. Jonassen, Systematic exploration of error sources in pyrosequencing flowgram data. *Bioinformatics*. **27**, 304–309 (2011).

311. J. G. Caporaso *et al.*, QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*. **7**, 335–336 (2010).

312. R. Schmieder, R. Edwards, Quality control and preprocessing of metagenomic datasets. *Bioinformatics*. **27**, 863–864 (2011).

313. R. Schmieder, R. Edwards, Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS One*. **6** (2011), doi:10.1371/journal.pone.0017288.

314. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. **26**, 589–595 (2010).

315. R. C. Edgar, Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. **26**, 2460–2461 (2010).

316. T. Z. DeSantis *et al.*, Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069–5072 (2006).

317. D. McDonald *et al.*, The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *Gigascience*. **1**, 7 (2012).

318. J. G. Caporaso *et al.*, PyNAST: A flexible tool for aligning sequences to a template alignment. *Bioinformatics*. **26**, 266–267 (2010).

319. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One*. **5** (2010), doi:10.1371/journal.pone.0009490.

320. Babraham Institute, FastQC, (available at https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

321. Illumina Inc, Illumina Adaptor Sequences, (available at https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/experiment-design/illumina-adapter-sequences_1000000002694-01.pdf).

322. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. **17**, 10 (2011).

323. B. Liu *et al.*, COPE: An accurate k-mer-based pair-end reads connection tool to facilitate genome assembly. *Bioinformatics*. **28**, 2870–2874 (2012).

324. Hannon Laboratory Cold Spring Harbor, FASTX-Toolkit, (available at http://hannonlab.cshl.edu/fastx_toolkit/).

325. NCBI, Gene2xml, (available at https://www.ncbi.nlm.nih.gov/IEB/ToolBox/C_DOC/lxr/source/).

326. Y. Zhou, Y. Liang, K. H. Lynch, J. J. Dennis, D. S. Wishart, PHAST: A Fast Phage Search Tool.

Nucleic Acids Res. **39**, 347–352 (2011).

327. R Development Core Team, R: A language and environment for statistical computing. (2008), (available at http://www.r-project.org).

328. R. García-López, J. F. Vázquez-Castellanos, A. Moya, Fragmentation and Coverage Variation in Viral Metagenome Assemblies, and Their Effect in Diversity Calculations. *Front. Bioeng. Biotechnol.* **3**, 1–15 (2015).

329. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. **9**, 357–9 (2012).

330. A. Chao, Non-parametric estimation of the classes in a population. *Scand. J. Stat.* **11**, 265–270 (1984).

331. C. E. Shannon, The mathematical theory of communication. 1963. *MD. Comput.* **14**, 306–17 (1948).

332. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc.* **57** (1995), pp. 289–300.

333. J. N. Paulson, O. C. Stine, H. C. Bravo, M. Pop, Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods*. **10**, 1200–1202 (2013).

334. J. R. Bray, J. T. Curtis, An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecol. Monogr.* **27**, 325–349 (1957).

335. P. Jaccard, Etude de la distribution florale dans une portion des Alpes et du Jura. *Bull. la Soc. Vaudoise des Sci. Nat.* **37**, 547–579 (1901).

336. B. H. McArdle, M. J. Anderson, Fitting multivariate models to community data: a comment based on distance-based redundancy analysis. *Ecology*. **82**, 290–297 (2001).

337. Y. Vázquez-Baeza, M. Pirrung, A. Gonzalez, R. Knight, EMPeror: a tool for visualizing high-throughput microbial community data. *Gigascience*. **2**, 16 (2013).

338. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

339. N. Segata *et al.*, Metagenomic biomarker discovery and explanation. *Genome Biol.* **12**, R60 (2011).

340. A. J. Rosnah Binti Zain, Thomas George Kallarakkal, Anand Ramanathan, Jin Kim, WM Tilakaratne, Takashi Takata, Saman Warnakulasuriya, Vinay Kumar Hazarey, Alison Rich, Haizal Mohd Hussaini, Exophytic Verrucous Hyperplasia of the Oral Cavity – Application of Standardized Criteria for Diagnosis from a Consensus Report. *Asian Pacific J. Cancer Prev.* **17**, 4491–4501 (2016).

341. L. Bagan *et al.*, Prevalence of salivary Epstein-Barr virus in potentially malignant oral disorders and oral squamous cell carcinoma. *Med. Oral Patol. Oral Cir. Bucal*. **21**, e157–e160 (2016).

342. S. Gupta, S. Gupta, Role of human papillomavirus in oral squamous cell carcinoma and oral potentially malignant disorders: A review of the literature. *Indian J Dent*. **6**, 91–98 (2015).

343. D. L. Capella, J. M. Gonçalves, A. A. A. Abrantes, L. J. Grando, F. I. Daniel, Proliferative verrucous leukoplakia: Diagnosis, management and current advances. *Braz. J. Otorhinolaryngol.*, 1–9 (2016).

344. M. A. L. da Silva *et al.*, A comparison of four DNA extraction protocols for the analysis of urine from patients with visceral leishmaniasis. *Rev. Soc. Bras. Med. Trop.* **47**, 193–197 (2014).

345. B. G. Fanson, P. Osmack, A. M. Di Bisceglie, A comparison between the phenol-chloroform method of RNA extraction and the QIAamp viral RNA kit in the extraction of hepatitis C and GB virus-C/hepatitis G viral RNA from serum. *J. Virol. Methods*. **89**, 23–27 (2000).

346. D. C. Rio, M. Ares, G. J. Hannon, T. W. Nilsen, Purification of RNA using TRIzol (TRI Reagent). *Cold Spring Harb. Protoc.* **5**, 2010–2013 (2010).

347. M. Džunková *et al.*, Direct squencing from the minimal number of DNA molecules needed to fill a 454 picotiterplate. *PLoS One*. **9**, e97379 (2014).

348. R. Marine *et al.*, Caught in the middle with multiple displacement amplification: the myth of pooling for avoiding multiple displacement amplification bias in a metagenome. *Microbiome*. **2**, 3 (2014).

349. J. R. Kugelman *et al.*, Error baseline rates of five sample preparation methods used to characterize RNA virus populations. *PLoS One*. **12**, 1–13 (2017).

350. B. E. Pickett *et al.*, ViPR: An open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.* **40**, 593–598 (2012).

351. R. Liechti *et al.*, *Database (Oxford).*, in press, doi:10.1093/database/baq004.

352. E. Foley B, Leitner T, Apetrei C, Hahn B, Mizrachi I, Mullins J, Rambaut A, Wolinsky S, and Korber B, HIV Sequence Compendium 2013. *Theor. Biol. Biophys. Group, Los Alamos Natl. Lab. NM, LA-UR 13-26007.* (2013) (available at http://www.hiv.lanl.gov/content/sequence/HIV/mainpage.html).

353. P. Masson *et al.*, ViralZone: Recent updates to the virus knowledge resource. *Nucleic Acids Res.* **41**, 579–583 (2013).

354. M. Stano, G. Beke, L. Klucar, ViruSITE - Integrated database for viral genomics. *Database*. **2016**, 1–6 (2016).

355. G. Cochrane, I. Karsch-Mizrachi, T. Takagi, The international nucleotide sequence database collaboration. *Nucleic Acids Res.* **44**, D48–D50 (2016).

356. D. Huson, A. Auch, J. Qi, S. Schuster, MEGAN analysis of metagenome data. *Genome Res.* **17**, 377–386 (2007).

357. P. Menzel, K. L. Ng, A. Krogh, Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* **7**, 11257 (2016).

358. J. Wallis, S. Cho, R. Diecidue, Propionibacterium Acnes: An Opportunistic Oral Pathogen. *J. Oral Maxillofac. Surg.* **69**, e5 (2011).

359. C. V. Hughes, P. E. Kolenbrander, R. N. Andersen, L. V. H. Moore, Coaggregation properties of human oral Veillonella spp.: Relationship to colonization site and oral ecology. *Appl. Environ. Microbiol.* **54**, 1957–1963 (1988).

360. A. Sadhu, R. Loewenstein, S. A. Klotz, Rothia dentocariosa endocarditis complicated by multiple cerebellar hemorrhages. *Diagn. Microbiol. Infect. Dis.* **53**, 239–240 (2005).

361. M. D. Stenglein *et al.*, Widespread recombination, reassortment, and transmission of unbalanced compound viral genotypes in natural arenavirus infections. *PLoS Pathog.* **11**, e1004900 (2015).

362. S. R. Radoshitzky *et al.*, Past, present, and future of arenavirus taxonomy. *Arch. Virol.* **160**, 1851–1874 (2015).

363. M. A. Ghannoum *et al.*, Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. *PLoS Pathog.* **6** (2010), doi:10.1371/journal.ppat.1000713.

364. H. Xu, A. Dongari-Bagtzoglou, Shaping the oral mycobiota: Interactions of opportunistic fungi with oral bacteria and the host. *Curr. Opin. Microbiol.* **26**, 65–70 (2015).

365. G. B. Huffnagle, M. C. Noverr, The emerging world of the fungal microbiome. *Trends Microbiol.* **21**, 334–341 (2013).

366. A. Edlund, T. M. Santiago-Rodriguez, T. K. Boehm, D. T. Pride, Bacteriophage and their potential roles in the human oral cavity. *J Oral Microbiol.* **7**, 27423 (2015).

367. M. Avila, D. M. Ojcius, Ö. Yilmaz, The Oral Microbiota: Living with a Permanent Guest. *DNA Cell Biol.* **28**, 405–411 (2009).

368. T. Elbeaino, R. A. Kubaa, H. T. Tuzlali, M. Digiaro, Pittosporum cryptic virus 1: genome sequence completion using next-generation sequencing. *Arch. Virol.* **161**, 2039–2042 (2016).

369. E. K. Binga, R. S. Lasken, J. D. Neufeld, Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. *ISME J.* **2**, 233–241 (2008).

370. S. W. Polson, S. W. Wilhelm, K. E. Wommack, Unraveling the viral tapestry (from inside the capsid out). *ISME J.* **5**, 165–168 (2011).

371. M. L. Gillison *et al.*, Prevalence of Oral HPV Infection in the United States, 2009-2010. *Jama*. **307**, 693 (2012).

372. R. Bouziat *et al.*, Reovirus infection triggers inflammatory responses to dietary antigens and development of celiac disease. *Science.* **356**, 44–50 (2017).

373. D. Paez-Espino *et al.*, IMG/VR: A database of cultured and uncultured DNA viruses and retroviruses. *Nucleic Acids Res.* **45**, D457–D465 (2017).

374. A. Reyes, N. P. Semenkovich, K. Whiteson, F. Rohwer, J. I. Gordon, Going viral: next-generation sequencing applied to phage populations in the human gut. *Nat. Rev. Microbiol.* **10**, 607–617 (2012).

375. L. K. Ursell, J. L. Metcalf, L. W. Parfrey, R. Knight, Defining the human microbiome. *Nutr. Rev.* **70** (2012), doi:10.1111/j.1753-4887.2012.00493.x.

376. E. Scarpellini *et al.*, The human gut microbiota and virome: Potential therapeutic implications. *Dig. Liver Dis.* **47**, 1007–1012 (2015).

377. L. Margulis, R. Fester, Eds., *Symbiosis as a Source of Evolutionary Innovation* (MIT Press, London, 1991).

378. M. P. Weekes *et al.*, Quantitative temporal viromics: An approach to investigate host-pathogen interaction. *Cell.* **157**, 1460–1472 (2014).

379. L. L. Li, A. Norman, L. H. Hansen, S. J. Sørensen, Metamobilomics - expanding our knowledge on the pool of plasmid encoded traits in natural environments using high-throughput sequencing. *Clin. Microbiol. Infect.* **18**, 5–7 (2012).